March 2017

Not Just Noise: Individual Differences in Cognitive Ability and Response Bias

Tina Chen
University of Massachusetts - Amherst

NOT JUST NOISE: INDIVIDUAL DIFFERENCES IN COGNITIVE ABILITY AND RESPONSE BIAS

A Dissertation Presented

by

TINA CHEN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2017

Department of Psychological and Brain Sciences

NOT JUST NOISE: INDIVIDUAL DIFFERENCES IN COGNITIVE ABILITY AND RESPONSE BIAS

A Dissertation Presented

by

TINA CHEN

Approved as to style and content by:

____
Caren Rotello, Chair

____
Jeffrey Starns, Member

____
Rebecca Ready, Member

____
Craig Wells, Member

____
Harold Grotevant, Department Head
Department of Psychological and Brain Sciences

DEDICATION

For my parents. I love you.

ACKNOWLEDGMENTS

First, thank you to my amazing advisor Caren Rotello. Your unwavering support, gentle guidance, and academic model have been instrumental in my career path. Thank you for always having an open door and being willing to stop whatever you are doing to help me, whether that is answering a quick administrative question or puzzling over data and talking me down from a panic. Your mentoring has been invaluable, allowing me to explore both research and career opportunities in a safe environment, and I don’t know if I can do enough to demonstrate my appreciation.

Thank you also to my dissertation committee: Jeff Starns, Becky Ready, and Craig Wells. I appreciate the thoughtful questions and feedback you’ve provided throughout this process; your expertise has been so helpful. I’d like to thank Aline Sayer for your support and help as well. And thank you to Ben Zobel for your E-Prime knowledge and to Colin Quirk and Brian Dillon for your JavaScript knowledge: programming my dissertation without your help would have been so much more painful.

A huge thanks to friends here who have been so essential to this journey: Andrea Cataldo, Will Hopper, and Lap Keung (who indulged my every question about things I should have already known); Merika Wilson and Junha Chang (awesome officemates); Lisa Fiorenzo and Sam Rhodes (who provided wide-ranging support); my GWIS peer mentoring group (for the advice and accountability, especially during the summers!); and my support at MERCYhouse (for grounding me in something other than academia).

And finally, thank you to my family. Thank you, Mom and Dad, for all of your faith in me and my ability. Thank to Peter and Vanessa for providing an occasional retreat. Thank you, Mary, for lamenting with me and celebrating with me.

ABSTRACT

NOT JUST NOISE: INDIVIDUAL DIFFERENCES IN COGNITIVE ABILITY AND RESPONSE BIAS

FEBRUARY 2017

TINA CHEN, B.S., UNIVERSITY OF OKLAHOMA

M.A., UNIVERSITY OF MISSOURI COLUMBIA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Caren M. Rotello

Response bias is a component of decision making that can be defined as the general willingness to respond a certain way. For example, in recognition memory, one can have a response bias towards responding that a test item has been previously studied, or in reasoning, one can have a response bias towards responding that a conclusion is logically valid. However, not all individuals have the same response bias. Indeed, there is some evidence that response bias is a stable cognitive trait in memory that differs across individuals (Kantner & Lindsay, 2012, 2014). One predictor of this trait may be cognitive ability, since it appears to predict response bias in memory (Zhu et al., 2010) and in reasoning (e.g., Handley & Trippas, 2015). While memory and reasoning have similar decision making components and may be very related (Heit & Hayes, 2011; Heit, Rotello, & Hayes, 2012), this experiment is the first to test whether cognitive ability predicts response bias in both tasks. Experiment 1 showed that higher cognitive ability participants were more conservative than lower cognitive ability participants in reasoning, but not in memory. Experiment 2 showed that participants did generally follow task demands to shift their bias some, but this shift was not predicted by cognitive ability. This study shows that further research is needed to examine individual differences in response bias as one way to account for what has previously been treated as noise.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

RESPONSE BIAS

Introduction

Response bias, i.e., one’s general willingness to pick an option or make a behavioral response over another, affects all kinds of decision making. For example, when deciding if you should wave back at the person who is waving at you, there is some influence of the strength of recognition for that person, but there is also a level of general willingness to wave at people. Some people might wave even if they have only a little bit of evidence that they know the person, perhaps because they weigh the cost of waving lower than the benefit of seeming friendly. Another example of response bias in decision making might be when deciding if your friend has made a sound, logical argument. There is some degree of the strength of the argument, but there may also be a level of general willingness to believe the friend has made a good argument; perhaps the friend is a smart or reasonable person. Notably, this general willingness to make a particular behavioral response is *independent* of any underlying evidence or strength for that response. Still, response bias does affect the behavior.

Response bias is important in behavioral cognitive psychology because while the *processes* are cognitive, the *data* of interest are behavioral: inferences about mental processes are made based on the behavioral response, which is of course influenced by response bias. Response bias has been thoroughly described in the context of recognition memory (e.g., Macmillan & Creelman, 1990), so a discussion of it in those terms follows. In the context of a recognition study, participants may have a general inclination to respond that a test item was previously studied, i.e., "old," or that it was not previously studied, i.e., "new." Response bias in memory research is generally operationalized as one's willingness to respond "old" to an item; a *liberal response bias* indicates a *greater* willingness to respond "old" and a *conservative response bias* indicates a *lesser* willingness to respond "old." This response bias is independent of the retrieval process or the evaluation of the evidence of memory strength, i.e., the ability to discriminate between old (target) and new (lure) items (Macmillan & Creelman, 1990). That is, both discriminability and response bias influence the behavioral response of indicating an item is “old,” but they are not the same process. A similar response bias occurs in reasoning tasks as well; for example, participants may have a greater or lesser willingness to respond that a conclusion of a logic syllogism is valid or invalid.

Response bias can be manipulated by the researcher. For example, a base rate manipulation influences bias by altering the proportion of old (or valid) items on the test; with a higher proportion of old items, one should have a more liberal response bias. Alternatively, response bias can be affected by rewarding certain responses more heavily than others in a payoff manipulation. While response bias is important in any cognitive task that requires decision making or discrimination, like categorization or perception, my focus will be on response bias particularly in memory and reasoning, and more specifically, on recognition memory tasks and a syllogistic reasoning task.

The importance of examining individual differences in response bias

Response bias is a decision making factor that is often different from the behavior of interest (one that reflects a certain cognitive process), but it can influence our understanding of accuracy or discriminability, which is often the process of interest. Because certain measures conflate bias and discriminability (see Rotello & Macmillan, 2008), it is important to account for each construct in measurement.

Accurately measuring response bias with quantitative models. While multiple models have been developed to separate these constructs (e.g., Pazzaglia, Dube, & Rotello, 2013), there is ample evidence in support of signal detection models (e.g., Banks, 1970; Macmillan & Creelman, 2005), especially for the memory and reasoning tasks used in the current study (e.g., Dube, Rotello, & Heit, 2010; Pazzaglia et al., 2013). Therefore, the most relevant model for this study is the signal detection model, and a discussion of that model follows.

Response bias in the signal detection model. The simplest signal detection model (see Figure 1A) posits that items in memory vary in strength along an axis of strength of evidence or familiarity (Macmillan & Creelman, 2005). Items that were studied in the list (targets, represented by the Normal distribution in the solid line) have higher memory strength and are shifted to the right in comparison to items that were not recently studied (lures, represented by the Normal distribution in the dotted line). Because these two distributions have the same variance, this model is referred to as the equal-variance signal detection model (EVSD). According to this model, participants who are asked to respond if an item is old or new will place a decision criterion c somewhere along the strength axis. So items are rated as “old” when the memory strength for the item exceeds the decision criterion. In other words, only items that have high enough memory strength will be identified as “old.” Criteria farther to the right in the figure are more conservative: lesser willingness to respond “old” unless the evidence is very strong. Conversely, criteria farther to the left are more liberal: greater willingness to respond “old” even when the evidence is not very strong.

If participants are asked to rate the confidence of their response on a 6-point scale ranging from "very sure new" to "very sure old," the same logic applies but includes more criteria to account for the different responses in between. Items are given the "very sure old" rating when the memory strength for that item exceeds that criterion, or as in Figure 1A, falls to the right of the rightmost vertical line c_5 when the criteria are numbered in order from left to right. Items are given the next highest confidence "sure old" rating when memory strength falls between c_4 and c_5, etc.

There is usually some overlap in the target and lure distributions such that depending on the placement of the criteria, some lures may be falsely recognized as old items. Correctly identifying a target as an old item is referred to as a "hit" and incorrectly identifying a lure as an old item is referred to as a "false alarm"; incorrectly identifying a target as new is a "miss" and correctly identifying a lure as new is a "correct rejection." If memory for the studied items is very good, the distributions are farther apart and the measure of sensitivity, d', i.e., the distance between the means of the distributions measured in standard deviation units, is larger: greater discriminability between old and new items. This discriminability is distinct from the decision criteria c_1 through c_5 which could be placed at different memory strengths regardless of the distance between the distributions.

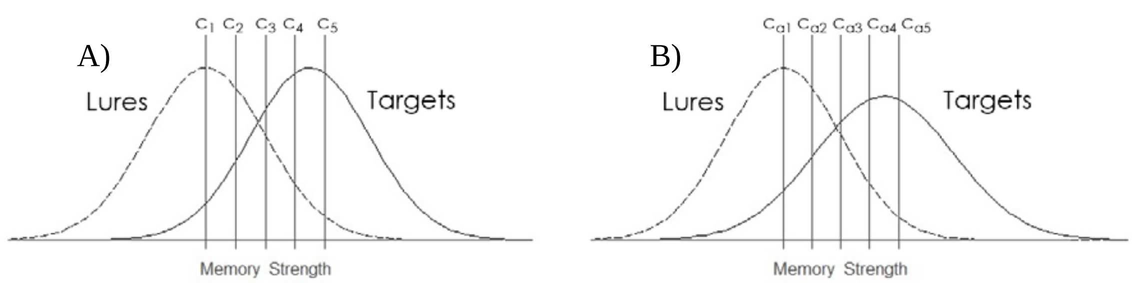


Figure 1. A) The equal-variance signal detection (EVSD) model with five decision criteria, represented as vertical lines, placed. The target distribution is represented with the solid line and the lure distribution is represented with the dotted line. B) The unequal-variance signal detection (UVSD) model also with five decision criteria placed. The target distribution is represented with the solid line and the lure distribution is represented with the dotted line.

While Figure 1 presents theoretical distributions, one way to depict the empirical data is in the form of receiver operating characteristic (ROC) curves, by plotting the ratings for targets against the ratings for lures. Figure 2 shows an ROC with a strong condition (based on the theoretical EVSD model in Figure 1A) in solid points and a weak condition in open points. Specifically, with a 6-point scale, the first (left-most) point for the condition plots the rate of responding “very sure old” to targets (y-coordinate) against the rate of responding “very sure old” to lures (x-coordinate). To map onto the theoretical distribution in Figure 1A, this is the proportion of the target distribution to the right of the rightmost criterion plotted against the proportion of the lure distribution to the right of that same criterion. The second point plots the rate of responding “very sure old” *and* “sure old” to targets against the rate of responding thusly to lures; i.e., it plots the cumulative sum of the hit rates and false alarm rates. In Figure 1, this is the proportion of the target and lure distributions to the right of the second rightmost criterion. Note that this second point includes the data from the first point, so it cannot fall to the left or below the previous one in Figure 2. But, if there were, for example, no “sure old” responses to a lure, and therefore no additional false alarms to be included in this point, the second point would fall immediately above the first point (by however many more hits from “sure old” responses to targets would be included). The third point in this ROC, marked with an x in Figure 2, indicates the cumulative sum of all of the “old” responses, providing an indicator of a binary decision between “old”/”new”, i.e., the more standard definition of a hit rate and false alarm rate. By plotting the rest of the cumulative rates, a curve emerges; there are only 5 visible points because the 6^{th} point, the cumulative rate of using the whole scale, must fall on the upper right hand corner (1, 1). In an ROC, points to the left indicate more conservative responses, or lower rates of responding “old” in this case, and points to the right indicate more liberal response, or higher rates of responding “old”.

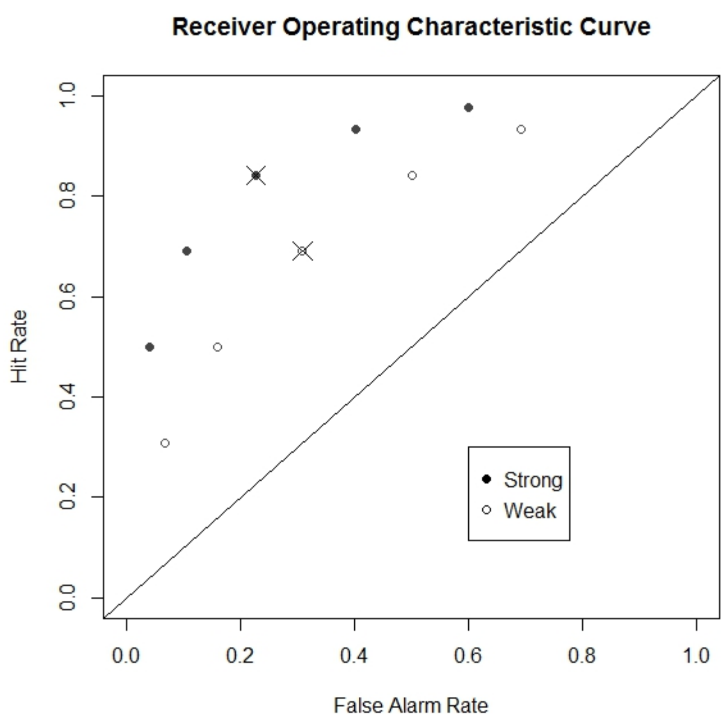


Figure 2. Idealized receiver operating characteristic curves (ROCs) for strong (from the EVSD model in Figure 1) and weak conditions using a 6-point confidence scale. The x through the third point in the ROC indicates the typically discussed hit and false alarm rate (“old” responses to old items and “old” responses to new items, regardless of confidence).

Also depicted in the ROC is an indication of discriminability, *d′*, between targets and lures. Curves that are closer to the top left corner indicate greater *d′*; curves that fall along the major diagonal (bottom left to top right corners) indicate *d′* of 0. In Figure 2, the strong condition is shifted towards the upper left corner in comparison to the weak condition, reflecting greater discriminability in the strong condition. Indeed, d' and the criteria can be found by fitting the signal detection model to the empirical data, finding the best curve and criteria that fit the data, with some restrictions. For one, the end points of this curve are at (0, 0) and (1, 1), the respective cumulative sums, though they are not explicitly plotted on the ROC. Another restriction of the curve is that the ROC is symmetrical along the minor diagonal (top left to bottom right corners) because these data assume that the target and lure distributions have the same variance.

Despite this, there is ample empirical evidence that the simple model presented in Figure 1A is not descriptive of the data, and an unequal-variance signal detection (UVSD) model, such as the one presented in Figure 1B, is more appropriate (e.g., Mickes, Wixted, & Wais, 2007; Ratcliff, Sheu, & Gronlund, 1992). This model is very similar to the EVSD model except that the target distribution has more variability than the lure distribution. The measures of d' and c are no longer appropriate since those are measured in terms of the common standard deviation unit of the target and lure distributions, which no longer applies. Instead, the measures of d_a, and c_a are the more general forms of the sensitivity and response bias measures and do not assume that the target and lure distributions have equal variance. The formula for d_a, is as follows:

$$d_a = \left(\frac{2}{1+s^2}\right)^{\frac{1}{2}} [z(H) - sz(F)] \quad \text{(Eq. 1)}$$

where s, the slope, is the ratio of the lure distribution standard deviation to target distribution standard deviation, $z(H)$ is the z-score transformation of the proportion of "old" responses to old items (i.e., z-transformed hit rate) and $z(F)$ is the z-score transformation of the proportion of "old" responses to new items (i.e., z-transformed false alarm rate) (Macmillan & Creelman, 2005). If *s* is 1, i.e., it is an equal variance case, the formula is just *z(H) - z(F)*, or the formula for *d′*. The formula for c_a is as follows:

$$c_a = \left(\frac{-\sqrt{2}s}{(1+s^2)^{\frac{1}{2}}(1+s)}\right)[z(H) + z(F)] \qquad \text{(Eq. 2)}$$

where the same variable definitions as above apply. If *s* is 1, i.e., it is an equal variance case, the formula is just *-1/2*[z(H) + z(F)]*, or the formula for *c*. With multiple responses (as in a confidence ratings design), multiple c_a could be calculated for the different responses.

Relative and Absolute Response Bias Measures and Optimal Response Bias

It should be noted that *c* and c_a measure *relative* response bias, i.e., the distance of the observed criterion location to the halfway point between the two distributions (e.g., Jones, Moore, Shub, & Amitay, 2015). In the EVSD context, *c* = 0 is located *d′*/2 standard deviation units to the right of the mean of the lure distribution, that is, where the distributions intersect. In this context, *c* = 0 is also the optimal criterion location in that it minimizes error: at this location, false alarms and misses are minimized. Similarly, in the UVSD context, $c_a = 0$ is located $d_a/2$ standard deviation units (roughly, an average of the standard deviations of each distribution) to the right of the mean of the lure distribution. Note that this is *not* equal to the intersection of the distributions and also *not* where false alarms and misses are minimized because that point is determined by the inequality of the variances. The response bias measures *c* and c_a are relative to the midpoint of the two distributions; in the case of EVSD, that midpoint is also an optimal criterion location.

Another measure of response bias would be *absolute* response bias, i.e., the location along the x-axis centered at the mean of the lure distribution (also the 0 point used for *d′*). Absolute response bias can be measured with λ, calculated by *-z(F)*. Using the z-transformed false alarm rate still assumes a continuous distribution (a Normal distribution), whereas using just the false alarm rate does not (as a proportion, it assumes a binomial distribution). Using the z-transformed false alarm rate also allows for easy conversion to *c* when *d′* is known: *c* is the difference between the observed criterion location λ and the reference point *d′*/2. Note that larger positive λ values, shifted farther to the right of the lure distribution, reflect a more conservative absolute response bias, i.e., fewer false alarms.

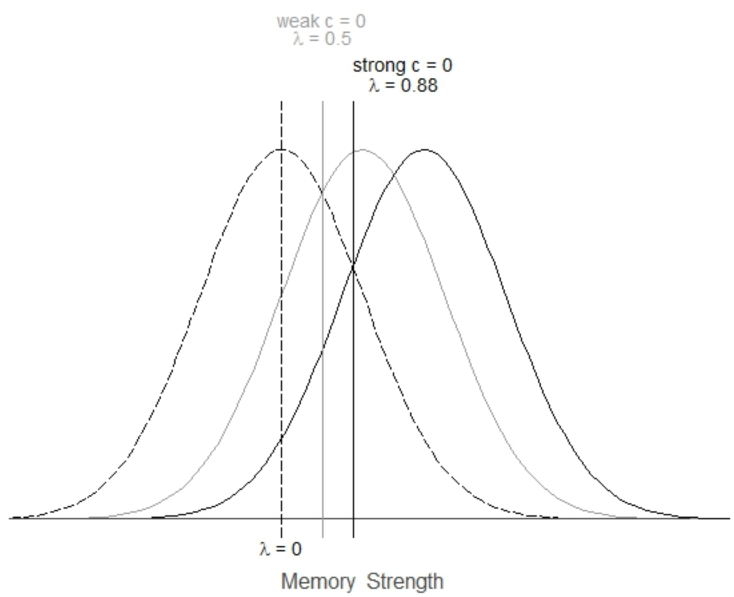


Figure 3. Zero points for relative and absolute response bias measures plotted on the strong (solid black line) and weak (solid gray line) target distributions and lure distribution (dashed black line) seen in Figure 2.

It is important to consider how different discriminability affects interpretation of these measures of response bias. Consider a situation where some targets have been repeated several times at study (strong discriminability) and some targets have been presented only once at study (weak discriminability) and assume an EVSD model for simplicity such as the one in Figure 3. In the strong condition, the target distribution is farther from the lure distribution, and so the halfway point is also farther. So in this case, $c = 0$ in the strong condition has shifted farther along the x-axis relative to where $c = 0$ is in the weak condition (λ of 0.88 vs. 0.5). Even if the criterion location does not shift along the x-axis across the two conditions, the value of c would still change because the point of reference (where $c = 0$) has changed: c reflects the distance from the optimal criterion location, but optimal has changed. However, λ is in relation to the lure distribution, which is centered at 0 on the x-axis, and so the interpretation of λ does *not* change with differences in discriminability: it still reflects the criterion location for that particular false alarm rate. Yet, because it does not take into consideration the hit rate, λ is an imperfect measure of response bias. Reasons to use relative vs. absolute response bias depend on the construct of interest. If the interest is in how people change response bias regardless of discriminability, then λ may be more appropriate. If the interest is in how people change response bias while accounting for differences in discriminability or how they change in relation to an ideal location such as $c = 0$, then relative response bias may be more appropriate.

Indeed, different strategies might be used by the participant to determine where to place decision criteria (Macmillan & Creelman, 2005). For example, in the waving example introduced earlier, perhaps a person wants to set an optimal decision rule that minimizes her “wrong” decisions about waving: she wants to avoid waving to people she does not know (reduce her false alarm rate) but also avoid *failing* to wave to people she *does* know (reduce her miss rate). In the context of a recognition study, if a participant wanted to minimize the overall error rate, the optimal decision criterion placement of the old/new boundary in the EVSD model would be where the heights of the distributions are the same, i.e., the two distributions crossover. Indeed, another measure of response bias is the likelihood ratio (β) which is the ratio of the height of the target distribution (the likelihood for the target distribution) to the height of the lure distribution (the likelihood for the lure distribution). Therefore, when the heights of those distributions are the same and so the proportions of correct "old" and "new" responses are maximized, the likelihood ratio is 1. This optimal likelihood ratio which maximizes accuracy is not exactly at 1 in the UVSD case, depending on the specific variances, but it is generally close to 1.

Some participants may see the costs and benefits of such responses differently. One might think it is better to be seen as friendly and wave to others, even if you might not know them. In the context of a recognition study, if a participant did not weigh correct "old" and "new" responses equally and wanted to maximize a *weighted* combination of the correct responses, the optimal decision criterion placement would be different; in other words, what is the *weighted* benefit of a hit vs. a correct rejection? This weighting could be internally motivated or externally motivated. If there were external motivations, such as in a payoff manipulation where experimenters may choose to reward hits heavily and penalize false alarms less heavily in order to shift criterion to a more liberal one, the optimal decision criterion placement would depend on the amount of reward/penalty in order to maximize the expected value (Macmillan & Creelman, 2005).The formula for the weighted likelihood ratio is as follows:

$$\mathrm{LR}(x) = \beta = \left(\frac{[R(correct\ rejection) - R(false\ alarm)]}{[R(hit) - R(miss)]}\right)\frac{p(S_1)}{p(S_2)} \qquad \text{(Eq. 3)}$$

where $p(S_1)$ refers to the probability of presenting a lure and $p(S_2)$ is the probability of presenting a target. Imagine a test with an equal proportion of targets and lures, and for simplicity, the researcher does not penalize for incorrect responses. If correct rejections are only rewarded with 1 cent and hits are rewarded with 10 cents, the likelihood ratio would be 1/10 or 0.1. A likelihood ratio less than 1 corresponds to a more liberal response bias; in this case, the decision criterion would be placed where the height of the target distribution is only 1/10 the height of the lure distribution, or much farther to the left on the strength axis. It is worth noting that if only the payoff is manipulated, participants cannot simultaneously maximize expected reward and accuracy because the decision rules are different in that situation (with a 10 cent reward for hits and a 1 cent reward for correct rejections, participants can either maximize expected value with likelihood ratio of 0.1 or maximize accuracy with likelihood ratio of 1 as mentioned earlier, but they cannot maximize both). Indeed, there is evidence that given the tension between those decision rules, participants will compromise between the two (e.g., Bohil & Maddox, 2001).

It seems reasonable to believe that these strategies could depend on the individual, and so it becomes very clear that merely looking at the proportion of correct "old" responses alone does not reflect underlying discriminability. That is to say, the proportion of “old” responses would be influenced by how liberal the response bias as well as the degree of discriminability. Thus, looking at certain accuracy measures like the proportion of “old” responses without accounting for response bias does not provide a precise or correct understanding of memory (Rotello & Macmillan, 2008).

The problem with averaging across individuals. In addition to needing to account for response bias by using such measures as c_a, it is important to account for the variability in response bias across individuals. Historically, models of cognition generally have been fit to aggregate data rather than to individual data, though increasingly, there is a call to avoid doing so because analyzing the aggregate can potentially obscure the nature of the processes of interest (e.g., Cohen, Sanborn, & Shiffrin, 2008). For example, Anderson and Tweney (1997) examined the shape of the forgetting curve in memory. Prior research aggregated across individuals and found they were best fit by a power function, which characterizes the rate of forgetting as decelerating over time. In their analysis, Anderson and Tweney found that individuals were best fit by an exponential function, which characterizes the rate of forgetting as constant across time. They attributed the different conclusion for the group data to artifacts from averaging across individuals; it is clear how such an artifact can impact cognitive theory (in this case, the nature of forgetting). More specifically to the topic at hand, it is important to also avoid analyzing the aggregate in response bias. If bias is different across individuals, averaging across them could obscure the nature of the processes of interest, such as how participants might shift their bias in response to a manipulation.

Some researchers have explicitly modeled individual differences in memory (Batchelder & Riefer, 1990; Malmberg, 2008) and individual differences in response bias in memory (Klauer & Kellen, 2010; Riefer, Hu, & Batchelder, 1994) and found that participants do indeed differ. However, while these individual differences are acknowledged to exist, the variance from individual differences is generally not of interest to cognitive psychologists because the underlying mental processes are assumed to be consistent across people, and most research is devoted to understanding these common processes.

In studies that do look at data at the individual level, the extra noise introduced by these individual differences may reduce the study's power to detect common, underlying processes of response bias. For example, Riefer and colleagues (1994) posited that the previous interpretation of source monitoring experiments could be confounded with response bias interpretations. While previous research had found that items generated by the participant are better remembered than those generated by another person, it could be the case that participants have a *bias* towards saying the other person generated the item rather than more *accurate* memory for items generated by themselves. On the aggregate level, Riefer et al. found that response bias differences across conditions could explain the findings of their first experiment. On an individual level, participants also differed in their response biases: participants with worse memory also had a greater response bias towards saying the other person generated the item. These kinds of individual differences in response bias, if considered to be noise, could reduce the power to detect overall response biases across the different conditions. While this example was a source memory recall study rather than a simple recognition memory study, and while Riefer et al. used a different measure of response bias than the ones described here, the fact remains that understanding individual differences can bolster our conceptualization, theories, and future research to fully capture the phenomena of interest, such as memory and reasoning.

Implications for experimental manipulations of bias. Not only is understanding individual differences in response bias important for theorizing about the constructs of memory and reasoning, but it is also important to improve our understanding of how individual differences in response bias may affect interpretation of our studies. Individual differences in response bias can interact with response bias manipulations in experimental designs. For example, Van Zandt (2000) varied bias in two ways: via the base rate (the proportion of targets and lures at test) and via payoff manipulations (monetary rewards and penalties for responding old vs. new). Participants rated their confidence that the test items were old or new. The bias manipulations led to different sensitivity as well as different ratios of standard deviations of the target and lure distributions within individuals. Van Zandt used this evidence to argue that ratings of confidence of items being old or new were not scaled directly from the memory evidence of those items, as is assumed by many models of recognition memory (Macmillan & Creelman, 1990; Ratcliff et al., 1992). It's possible that variability in criterion locations led to these differences in sensitivity and standard deviation ratios (Rotello & Macmillan, 2008; Treisman & Faulkner, 1984; Treisman & Williams, 1984); individual differences in response bias in using the confidence rating scale could have played a role. Particularly, individuals may vary in how they use the ratings scale and those differences may obscure the actual scaling from evidence to behavior. Indeed, Van Zandt (2000) noted that the effect of bias manipulations on slopes and intercepts varied across individuals: some individuals had large variability across the bias manipulations but some did not.

In sum, individual differences in response bias have not been thoroughly studied, in part because the process of response bias underlying cognition regardless of individual difference has been the primary focus. But even examining individual level data, commonly used measures of discriminability such as proportion correct are not independent of response bias. In this case then, not only are our conclusions about accuracy flawed, but individual differences in response bias would also influence these measures of accuracy, adding even more noise to the conclusions. It is important to use measures of discriminability and response bias that are independent and to ensure greater power from reducing noise by accounting for response bias differences. For example, averaging across confidence levels may obscure how participants may use confidence levels differently, which adds to the error term in statistical models; if we can predict or capture those differences in scale usage, we can reduce the unexplained variance. Additionally, participants may respond to bias manipulations in different ways, which reduces the power of those manipulations; if we can predict or capture those differences in response to bias manipulations, we can understand how effective those manipulations are.

Extant research on individual differences in response bias

Some research has been done on response bias in memory in special populations, generally finding more liberal response bias in: older adults in comparison to younger adults (Del Missier et al., 2013; Gilinsky & Judd, 1994; Huh, Kramer, Gazzaley, & Delis, 2006), older adults with dementia in comparison to age-matched controls (Woodard, Axelrod, Mordecai, & Shannon, 2004), older adults with Alzheimer's disease specifically in comparison to older adults with Huntington's disease (Brandt, Corwin, & Krafft, 1992), patients with schizophrenia in comparison to a control group (Brébion, Smith, Amador, Malaspina, & Gorman, 1998), and sub-clinical populations like people from a typical undergraduate student sample who were identified as highly anxious (Dowens & Calvo, 2003; Frenkel, Lamy, Algom, & Bar-Haim, 2009).

Despite these consistent findings that certain special populations have more liberal response bias, the study of individual differences in response bias within non-special populations has been inconsistent and sparse, perhaps due to a general disinterest in exploring what is considered to be a nuisance variable. Nonetheless, because of the potential issues that arise from individual differences, such as with interpretation of model parameters and bias manipulations, it is essential to understand what the extant literature says about individual differences in response bias in non-clinical populations for memory tasks and reasoning tasks.

Individual differences in response bias in memory

False memory. Almost all of the studies conducted thus far on individual differences in response bias in memory in non-clinical or special populations look at false memory rather than veridical memory. For example, in the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995), participants study lists of semantically related items. For example, participants may study a list that includes the words "nurse," "sick," "clinic," "health," "patient," and "medicine." After the study list, participants often remember the semantically related word "doctor" as having been presented in the list even though it did not appear (Roediger & McDermott, 1995). Participants, then, have false memory for the "critical lure," or the word that is likely to be falsely remembered based on the experimenter-constructed list. In another false memory paradigm, the misinformation paradigm (Loftus, 1975), participants are presented information about an event, such as a car accident. After this, they may be asked questions that include some false information about the event, such as, "How fast was the car going when it ran the stop sign?" (p. 563) even though there was no stop sign in the original event. When tested later, participants remember that a stop sign did appear in the original event (Loftus, 1975). Participants, then, have false memory for the misinformation.

While studies of false memory may not explicitly address response bias, it is reasonable to believe that there may be a connection between false memory and response bias. Indeed, liberal response bias on a standard recognition task weakly predicts increased false recall in DRM (Kantner & Lindsay, 2012) and liberal response bias was correlated with false cued-recall (Jonker, 2016). Given the relationship between liberal response criterion and false memory, understanding what individual differences predict false memory may provide direction for what individual differences predict liberal response bias.

The most promising of these is related to cognitive ability, defined here very broadly as the capacity and ability to store, process, and manipulate information either as in working memory (Baddeley & Hitch, 1974; Baddeley, 1996), executive functioning (Baddeley, 1996; Miyake et al., 2000) or intelligence (Cattell, 1994; Engle & Tuholski, 1999). While even the constructs within executive functioning alone (shifting, inhibition, and maintenance) are separable and not necessarily correlated with each other (Miyake et al., 2000), a broad definition of cognitive ability is a conservative approach to finding what may predict response bias because the literature thus far has used many different tasks. Their commonality appears to be a construct that includes the capacity to store, process, and manipulate information. For example, Zhu et al. (2010a) found that the cognitive factors of low intelligence, low perception, and low memory led to more false memories in the misinformation paradigm, and working memory capacity showed a negative relationship to false memory in the DRM paradigm. Watson, Bunting, Poole, and Conway (2005) proposed that working memory capacity influenced cognitive control and task goal maintenance because high working memory span individuals were able to integrate experimenter-provided warnings about false memory into a goal to reduce their false memory, but low span individuals were unable to do this. This goal maintenance and cognitive control could serve to minimize the influence of non-optimal response bias.

Even studies on personality predictors of false memory seem to indicate that cognitive ability is important: in another study, Zhu and colleagues (2010) found that several personality measures interacted with lower cognitive ability, as measured on several tests including working memory, to lead to more false memories in the misinformation effect: low fear of negative evaluation, low harm avoidance, high cooperativeness, high reward dependence, and high self-directedness. Need for Cognition, which is a measure of how much people tend to desire thinking or effortful cognition (Cacioppo & Petty, 1982), also predicts more susceptibility to false memory in DRM (Graham, 2007). Generally, cognition and cognitive ability appear to be important in predicting false memory.

While there is evidence that cognitive ability predicts false memory, which may be related to response bias, it’s important to also see if cognitive ability predicts response bias directly in veridical memory. Understanding false memory can lead to greater understanding of episodic memory generally, but individual differences may also affect veridical memory. An important first step is to see how individual differences manifest in the more common, general episodic memory paradigms in the absence of specific manipulations before understanding how they may interact with response bias manipulations.

Veridical memory. The extant research on individual differences in response bias approaches the research question in disparate ways. For example, Gillespie and Eysenck (1980) calculated response bias in a continuous recognition paradigm at two levels, strict response bias and lax response bias. They found that there was no difference between introverts and extraverts at the lax criterion, but that introverts were more conservative than extraverts at the strict criterion. This double stratification where individual differences (introverts and extraverts) in response bias (conservative and liberal) are examined within a particular range of bias (strict and lax criteria) is difficult to interpret and compare to more recent studies since it is an uncommon approach to measuring differences in response bias. Another approach, taken by Kantner and Lindsay (2012, 2014), was to use response bias in a standard recognition task to predict performance on other tasks. They found that conservative response bias in a standard recognition task weakly predicted fewer identifications in an eyewitness identification task (Kantner & Lindsay, 2014) and less false recall in a DRM paradigm (Kantner & Lindsay, 2012).

In one examination of individual differences in standard episodic recognition, Kantner and Lindsay (2012) examined which personality measures might predict response bias. They failed to find any relationship to personality measures like Need for Cognition, Behavior Inhibition and Avoidance Scales (BIS/BAS) which is a measure of how people tend to view approach/reward and avoidance/punishment (Carver & White, 1994), nor Maximizing and Regret which is a measure of how people try to maximize rewards or tend to just have a threshold of acceptability (Schwartz et al., 2002). Another example of the research into individual differences in response bias was conducted by

Aminoff and colleagues (2012) who specifically measured individual differences in response bias from a base rate manipulation. They looked at army officers and some graduate students, and presented words and faces to be remembered. At test, items were color coded as having a 70% probability of being old or as having a 30% probability of being old. Interestingly, this manipulation was done within list, though there is evidence that participants do not readily shift criterion within list (Stretch & Wixted, 1998). They found that some individuals were "high shifters" who employed widely variable response bias and others were "low shifters" who were more resistant to the bias manipulation. Aminoff and colleagues reported that shifting of response bias was mediated by *d′*. Shifting was also mediated by self-reported reliance on probabilistic cue information, and personality, specifically, the fun-seeking personality measure from BIS/BAS and negative affect from the Positive and Negative Affect Scale (PANAS) (Watson, Clark, & Tellegen, 1988). More recent work by Jonker (2016) demonstrated that individual differences in source memory also predict response bias. Participants were shown category-exemplar pairs (e.g., “FRUIT-apple”) at study, and then at test, cued with the category and asked to provide the exemplar. Then, participants were asked to rate their confidence that the item they provided was studied, which can be thought of as a memory for the source of the item, though conceptualized by Jonker primarily as a metamemory task. The accuracy of their source memory predicted response bias: participants who had lower source memory were more conservative. This runs counter to the other research that generally indicates lower cognition and cognitive ability leads to more liberal response bias, but it does provide further evidence for the relevance of those predictors in response bias.

Individual differences in response bias in reasoning

Though memory and reasoning are usually considered to be different domains, there is good reason to believe that they are related and that individual difference findings in one field could generalize to the other. For example, Heit and Hayes (2011) conducted eight experiments using the same stimuli for recognition memory and inductive reasoning tasks and found that performance on these tasks was very highly correlated across items. They argued that reasoning and recognition memory include an aspect of comparing novel and familiar stimuli. Individual differences in response bias would presumably affect both reasoning and memory, as both include these decision processes.

Indeed, Heit, Rotello, & Hayes (2012) argued that memory and reasoning are analogous and dependent on common mechanisms. Though they also presented data that showed that participants were significantly more liberal in a reasoning task than in the memory task, there could still be an effect of individual differences; this was not specifically explored in that study.

While the authors of studies of individual differences in reasoning rarely explicitly discuss the tasks' relationships to response bias in particular (Stupple, Ball, & Ellis, 2013; Stupple, Ball, Evans, & Kamal-Smith, 2011), there is reason to believe that even some standard reasoning tasks may be more related to response bias than at first glance. In particular, the belief bias effect warrants closer examination as an indicator of response bias. The belief bias effect occurs when participants are more likely to rate the conclusion of a syllogism as logically valid if the conclusion is believable, regardless of whether it follows logically. For example, consider whether the conclusion of this syllogism is logically valid:

If some doctors are surgeons,

And no women are surgeons,

*Then, some women are not doctors. (Gilinsky & Judd, 1994, pg. 360)

Though this is actually an invalid conclusion based on the premises presented, participants are more likely to rate this conclusion valid because that sentence coincides with prior knowledge and is believable: there are indeed some women who are not doctors. Historically, this has been described as an effect on reasoning accuracy (that participants are less accurate when the conclusion is believable than when the conclusion is unbelievable), but there is reason to believe the commonly used measure of accuracy is inappropriate and conflates discriminability and response bias (Dube et al., 2010). Dube, and colleagues (2010) found that when using the appropriate measures, which are the signal detection measures mentioned earlier, the belief bias effect is shown to be a response bias rather than a true accuracy effect. Believable conclusions shift participants' response biases towards "valid." Accordingly, individual differences in belief bias performance should approximate the individual differences in response bias in reasoning.

Generally, the individual differences in belief bias are consistent across experiments and with the previously discussed predictors of individual differences in response bias. Namely, working memory capacity and cognitive ability appear to predict performance in the belief bias paradigm. Participants high in cognitive ability are more accurate in the belief bias paradigm (Handley & Trippas, 2015; Newstead, Handley, Harley, Wright, & Farrelly, 2004; Sá, West, & Stanovich, 1999; Stanovich & West, 2008; Trippas, Handley, & Verde, 2013, 2014; Trippas, Verde, & Handley, 2014). Trippas and colleagues found that participants low in cognitive ability show a more liberal response bias to believable conclusions (Trippas, Verde, et al., 2014). This provides converging evidence that cognitive ability is important in response bias.

Thus, a common theme appears in reasoning as in memory: working memory/cognitive ability are important factors that explain individual differences in response bias. Indeed, a few studies show that response bias in memory tasks is relatively stable within individuals and could be thought of as a cognitive trait (Aminoff et al., 2012; Kantner & Lindsay, 2012, 2014). Given this stability and the consistent findings of the importance of cognitive ability, it seems that more should be done to establish the nature of the relationship.

Though some authors like Kantner and Lindsay (2012) have argued that individual differences in response bias are not worth studying if it is not a large effect, it is essential to explore individual differences in memory not only because doing so will allow for a fuller understanding of cognition, but also because the "extra noise" of individual differences reduces the power of finding the true underlying common processes. In particular, it is important to look at individual difference in response bias in memory because response bias affects behavioral measures of memory, "contamination" from these individual differences may affect parameter interpretation, and differences in response bias affect the assumptions and understanding of bias manipulations in experimental design. Understanding individual differences in response bias, and in particular those related to working memory and cognitive ability, can enrich understanding of cognitive processes and should be explored further. This research will provide some much-needed data that will allow for greater theoretical development in the future.

The overarching research purpose for this study is to take those steps towards understanding individual differences in response bias by examining the role of cognitive ability in response bias. In particular, this study looks at cognitive ability, memory, and reasoning and also seeks to answer whether cognitive ability can predict the optimality of response bias. That is, does cognitive ability predict response bias in both reasoning and memory, and do higher cognitive ability individuals choose a more optimal response bias?

CHAPTER 2

EXPERIMENT 1

In order to establish whether individual differences in cognitive ability predict response bias in both memory and reasoning, several measures of cognitive ability, a standard recognition memory task, and a standard belief bias reasoning task were administered to both a large online sample and a laboratory sample for comparison. For both the memory and reasoning task, I predicted that increased difficulty, defined by conditions that have historically been thought to decrease discriminability, would lead to more liberal absolute response bias/more false alarms (Hirshman, 1995) and conversely, lower difficulty to more conservative absolute response bias. Furthermore, I predicted that cognitive ability would influence response bias: higher cognitive ability participants may be better able to maintain the implicit task goal of maximizing accuracy, and therefore adjust their response bias accordingly to the difficulty of the task.

Specifically, in the memory task, I predicted that participants would have a more *conservative absolute* response bias in the *stronger condition* than in the weaker condition because they would be able to easily reduce false alarms without negatively impacting hits (Hirshman, 1995). Hirshman found that participants were more conservative when items had been presented for 2 seconds in comparison to when items had been presented for only 400 ms, consistent with the hypothesis that participants generate the possible range of memory strength for strong, weak, and new items at test and place their criterion accounting for that range. While he found effects using relative response bias *c*, predicting such effects on *c* in the current study actually requires a strong assumption about the effect. When discriminability increases, $c = 0$ shifts along the axis towards the target distribution, i.e., more conservative in absolute space. So in order to see a significant effect of c becoming more conservative in this strong condition, the shift in absolute space must equal the already more conservative shift of $c = 0$ plus an *additional* shift relative to c. Though I did not manipulate strength via the rate of presentation in the current experiment, manipulating strength with repetition should lead to a similar result: *repeated* items will lead to *more conservative response bias*. Because I do not make as strong an assumption about how large the shift must be, I predict only that there will be a shift in absolute response bias.

Beyond the main effect of repetition on response bias, I predicted that *higher cognitive ability* participants would be even better at adjusting their criterion depending on the difficulty of the condition while maximizing accuracy: *more conservative with an easier condition*.

In the reasoning task, I predicted that participants would have more *liberal* response bias in the *believable* condition than in the unbelievable condition (Dube et al., 2010; Trippas, Handley, et al., 2014). Though this prediction was phrased in the opposite format as the prediction for the memory task, the trend is in the same direction. The believable condition in the belief bias task is assumed by several theories to be more difficult than the unbelievable condition, potentially because one must override the believability of the syllogism in order to accurately judge the validity (e.g., Oakhill, Johnson-Laird, & Garnham, 1989; Stupple et al., 2011). While the actual effect on discriminability may not be real (Dube et al., 2010) and so the “more difficult” label may not be accurate, the prediction of the effect on response bias is also a replication of Dube and colleagues.

Beyond the main effect of believability on response bias, I predicted that higher cognitive ability participants would be better at adjusting their criterion while maximizing accuracy; these participants would be able to respond appropriately to the task (Stanovich, West, & Toplak, 2010): more conservative with an "easier" condition. In other words, higher cognitive ability participants were predicted to be more conservative in the "easier" condition of the recognition task (stronger/more repetitions) and in the reasoning task (unbelievable conclusions) than lower cognitive ability participants because higher cognitive ability participants would be able to respond appropriately to the task in order to maximize accuracy. However, the amount of shifting was predicted to be different in the memory and reasoning tasks. Specifically, higher cognitive ability participants would be expected to shift their bias more in the memory task in order to appropriately adjust to the task. However, in the reasoning task, this bias shift is not actually appropriate; higher cognitive ability participants should not shift their bias as much to be so liberal in the believable conclusion. Thus, though the amount of shifting predicted might be different, the underlying reason for that shift (responding appropriately to the task) remains the same, and I hypothesized that higher cognitive ability participants would be better at this.

Method

Participants

Because most samples in cognitive research come from the college student population, 103 participants from the University of Massachusetts recruited through the SONA website for the Department of Psychological and Brain Sciences were initially run in this experiment. Of those participants, 63 were dropped due to researcher error in programming the conditions, the randomization, and the counterbalancing of stimuli, which left a remaining sample of 40 participants, which is a sample size consistent with previous research on individual differences in response bias (Kantner & Lindsay, 2012). Every participant received credit for their participation. The lab sample mostly served as comparison to the online sample to ensure there would be no major differences in sample and therefore generalizability.

In order to achieve sufficient power, a much larger sample was recruited online through Amazon.com's Mechanical Turk via TurkPrime, a website designed to interface more clearly and easily with Mechanical Turk. Only Mechanical Turk "workers" with an IP address in the U.S. were able to participate, and eligibility was further determined with a prescreening survey administered through Qualtrics (a survey site) which paid 5 cents for answering 5 questions; if participants were eligible for the study, they were provided with a link to the "bonus" study administered through IbexFarm (a JavaScript-based programming language and host site) that would pay $3 for the rest of the experiment, estimated to take one hour. This payment was estimated to be comparable to other similarly timed tasks on Mechanical Turk: not as low as other tasks, but not too high as to be coercively enticing.

The prescreening survey asked about age and whether the participant had taken the SAT or ACT and could report the year and their score. An age restriction of 18-30 years old was chosen so that the online sample would be more comparable to the lab sample; this restriction was in the initial description. If participants reported an age outside of the acceptable range, they were shown a message that indicated they were not eligible for the study. The other main eligibility determinant was self-reported SAT and/or ACT score and year, which was one of the measures of cognitive ability, discussed in more detail below. If participants reported scores and years that were not plausible (e.g., scoring higher than maximum 36 on the ACT or testing in 1992 when the oldest allowed participant would have been 6 years old), they were shown a message that indicated they were not eligible for the bonus but could still receive payment for the prescreening. In other words, the responses to the prescreening served as a gatekeeper to the bonus survey and bonus money. Of the 1,024 participants who completed a prescreening to determine eligibility, 452 participants were eligible. Of these, only 350 of those participants were eligible and completed the full experiment.

Participants were excluded from analysis based on several factors. Two exclusion criterion were selected to ensure that the individual differences measures were meaningful. Specifically, participants who reported an implausible SAT/ACT score that had gotten through the Qualtrics filters, such as an SAT score that was not divisible by 10 but within the plausible range, were excluded. This applied to 21 online participants but none of the lab participants. Similarly, participants who performed poorly on the Go/No-Go task (described below), that is, lower than d' of 0.5, were also excluded; this applied to an additional 19 online participants and 1 lab participant. There were also two exclusion criteria set for the behavioral responses. First, median reaction times in the Number-Letter task (described below) that were less than 800 ms were assumed to indicate a failure to engage seriously with the task, and likewise median RTs greater than 2000 ms were assumed to indicate that the participant was probably distracted. However, no participants were excluded based on median RTs. Finally, participants were excluded if their confidence ratings and decision accuracy levels were not systematically related. Higher confidence responses should be more accurate than lower confidence responses (e.g., Wixted, Mickes, Dunn, Clark, & Wells, 2016). With a cut-off of percent correct being at least 10% greater in the highest confidence responses than the lowest confidence responses, 99 online participants and 10 lab participants for the memory task were excluded, as well as 32 online participants and 8 lab participants for the reasoning task. The remaining 179 online participants and 21 lab participants were included in the analyses.

Cognitive Ability Measures. To measure cognitive ability, multiple commonly used measures were used, since various measures have been used across different studies. It was assumed that these measures would relate enough to be considered convergent measures of cognitive ability.

Standardized academic testing. The first measure was self-reported SAT or ACT scores and the year in which they took the test. This has been used in previous studies as a proxy for cognitive ability (Naemi, Beal, & Payne, 2009; Stanovich & West, 1998, 2008). As mentioned above, online participants were excluded if they did not self-report plausible scores or years; lab participants, who were all undergraduates, were also required to self-report their scores/years.

Executive control. Working memory capacity tasks such as the common OSPAN task (Unsworth, Heitz, Schrock, & Engle, 2005) can be particularly difficult to administer online due to the potential lack of understanding the instructions, the potential for cheating by using external aids instead of memory, and the difficult implementation of the task calibrations based on performance. While working memory capacity has been commonly used as a measure of cognitive ability, for practical reasons, I used tasks that are related to a component of working memory: executive control.

According to the commonly used working memory model proposed by Baddeley and Hitch (Baddeley & Hitch, 1974; Baddeley, 1996), the non-storage component of working memory is the central executive. Miyake and colleagues (2000) characterized executive functioning in three components: inhibition (of prepotent responses); shifting between mental information; and updating or adjusting monitored information. The current study measured inhibition and shifting in individuals as a measure of cognitive ability.

Inhibition. While working memory capacity could be difficult to measure online, a Go/No-Go task could be appropriate to use instead: Redick, Calvo, Gay, and Engle (2011) found that working memory capacity, as measured by complex span tasks including OSPAN, was highly related to performance on a conditional Go/No-Go task. In addition to the findings of Redick and colleagues, further reason to use this task to measure cognitive ability came from Kantner and Lindsay (2014) who found evidence that response bias was predicted by a standard Go/No-Go task. Accordingly, it seemed reasonable to use this task, especially if working memory capacity measures would not be practical for an online sample.

In the conditional Go/No-Go task (Redick et al., 2011) used in this study, participants saw a series of single letters presented on the screen for 300 ms each followed by a blank screen for 700 ms during which they could also respond. They were told to respond to X, then, Y, then X, etc., regardless of whatever letters were presented in between, including duplicate Xs or Ys. When they saw any letter other than X or Y, they were not supposed to respond (distractor trial), nor when there was a duplicate X or Y from when they last responded (lure trial). So if they saw the series “X, H, Y, Y,” they should have responded to the X (target trial), not responded to the H (distractor trial), responded to the Y (lure trial) because the previous instance of a target letter was the *other* letter, but not responded to the second Y (lure trial) because the previous instance of a target letter was the *same* letter. In each of three blocks, there were 80 target trials, 20 lure trials, and 100 distractor trials. Accuracy was calculated with *d′* based on hits from target trials and false alarms from lure trials. Performance on a practice with feedback was not included in the accuracy score.

Shifting. The third measure of cognitive ability was the Number-Letter task, a task of the shifting component of executive functioning (Miyake et al., 2000). In this task as implemented by Miyake and colleagues, participants saw a number-letter pair in the top half of the screen or the bottom half of the screen. For pairs in the top half, they were asked to indicate whether the number was odd or even, and for pairs in the bottom half, they were asked to indicate whether the letter was a vowel or a consonant. This was cued with an instruction 150 ms before seeing the pair, with the instruction and pair staying on the screen until the participant replied after which there was a 350 ms blank screen before the next trial. In each of two blocks, there were 24 no-shift trials (trials where the previous task was the same) and 24 shift trials (trials where the previous task was different). The cost in reaction times of shifting between the number and letter task in relation to answering two number (or two letter) tasks in a row indicated a general ability to shift tasks.

Cognitive tasks

Memory task. The memory task included two conditions: words presented one time (1x) and words presented three times (3x). There were two study-test lists for each condition, with each study list containing 40 words presented for one second at a time. The words were 4 to 7 letter nouns, ranging in frequency from 5 – 787, with a mean of 49.07, taken from the MRC psycholinguistic database (Coltheart, 1981). For each test list, there were 40 targets/old words and 40 lures/new words. Participants responded on a six-point confidence scale the extent to which they were sure that an item appeared or did not appear on the study list from "very sure new" to "very sure old".

Reasoning task. The reasoning task included 32 syllogisms with 8 each of: valid and believable, valid and not believable, invalid and believable, and valid and not believable conclusions. These syllogisms were taken from Experiment 2 of Dube et al. (2010) which minimized effects of atmosphere and figure. The semantic content was counterbalanced, and the middle term in the premises (the noun that was not in the conclusion) was a nonsense word in order to reduce the believability of the premises. Participants responded on a six-point confidence scale the extent to which they were sure that the conclusion of the syllogism is logically valid from "very sure invalid" to "very sure valid".

Results

Variable Measurement, Approach to Analyses, and Hypotheses

The dependent variables of interest were absolute criterion location and relative criterion location, response bias variability as measured by the standard deviation of the respective measures across the conditions in individuals, and the total magnitude of the shift as measured by the range of the respective measures for each individual.

It should also be noted that the manipulations here, especially the repetition of stimuli in the memory task, usually increase accuracy. Finding individual differences in accuracy would not be problematic for the current experiments, and indeed would provide a manipulation check. Still, the primary focus of the analyses was on response bias rather than accuracy.

Approach to Analyses and Hypotheses

Cognitive ability was based on a composite score of SAT/ACT test score (converted to percentiles based on the year taken, which was then averaged if the participant took both tests), Go/No-Go accuracy (*d′* calculated with hits on target trials and false alarms on lure trials), and Number-Letter shifting cost (calculated by subtracting reaction time on no shift trials from reaction time on shift trials). Each individual's score on each task was converted to a z-score; these scores were then averaged to form a composite score on which the median split was based.

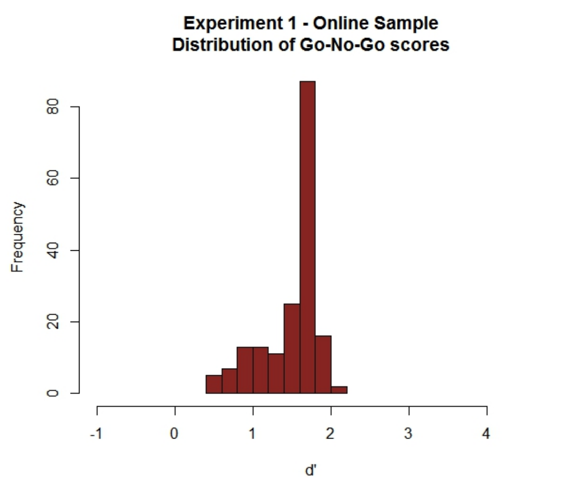
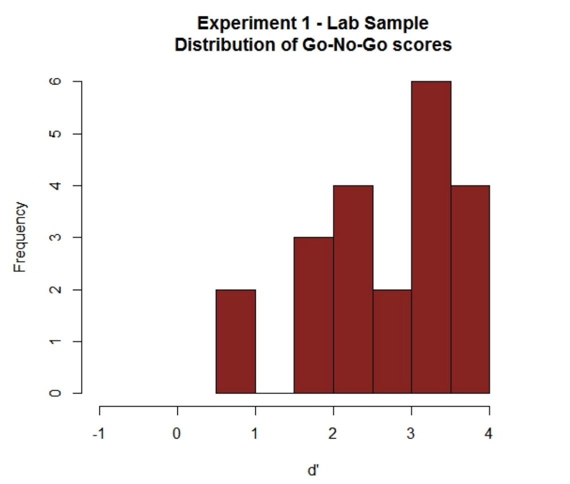
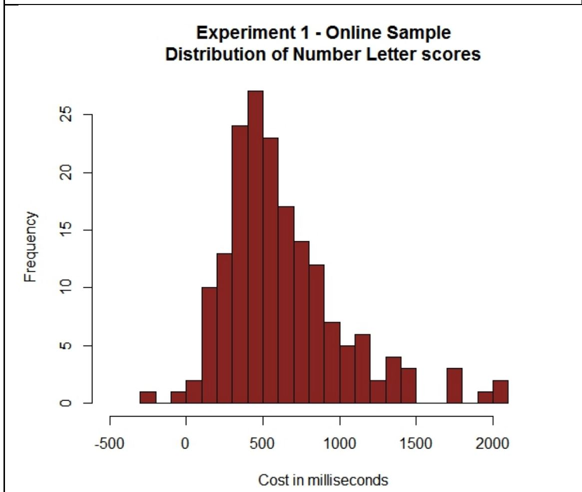
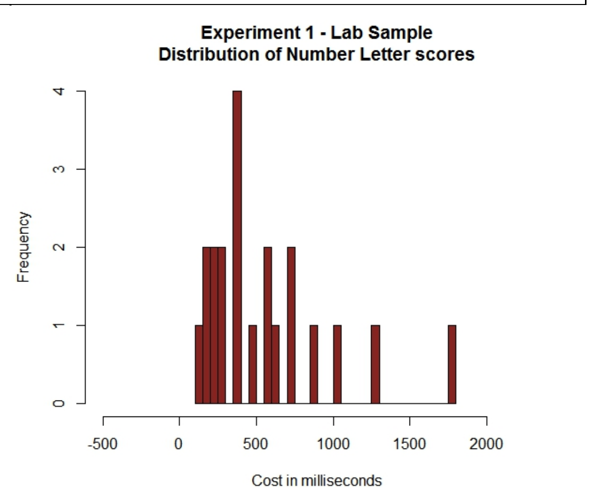
The composite score of cognitive ability was hypothesized to predict differences in response bias in the recognition and belief bias tasks. For each task and dependent variable, I conducted a 2-level hierarchical linear model (HLM) with condition at the level 1 and the individual at level 2, including cognitive ability as a predictor. Hierarchical linear models are linear models that account for the natural nesting that may occur in the data (Raudenbush & Bryk, 2002). For example, a mixed design study includes data from multiple time points that are nested within the individual, while individuals are nested in groups. Hierarchical linear modeling not only accounts for the related variance due to such nesting, but also allows for predictors at each level of nesting. It can also be used as another approach to ANCOVA, where a covariate would occur at the level of the individual, and the factors of interest at the relevant level. That is, while an ANCOVA would allow for both categorical predictors (condition) and continuous covariates (cognitive ability), this fails to capture the nested design: conditions are nested within individuals, which HLM allows to be explicitly modeled. Using hierarchical linear modeling allows for predictors at the individual level in a repeated measures design.

For the recognition memory task, I predicted that more repetition would lead to more conservative absolute response bias as Hirshman (1995) found, but more relevantly to my main research question, I predicted an interaction of cognitive ability and repetition such that *higher* cognitive ability participants would be particularly *more conservative* with *more repetition*. For the belief bias task. I predicted that believable conclusions would lead to more liberal response bias as Dube et al. (2010) found. I also predicted that there would be an interaction of cognitive ability and believability such that *lower* cognitive ability participants are particularly *more liberal* in responding that a syllogism is valid if it is *believable*. Thus, I predicted that individual differences in cognitive ability would predict absolute response bias in the memory and the reasoning tasks.

The results from the three measures of cognitive ability are presented first; comparing the performance in these tasks between the online and lab sample helps to determine the appropriateness of the online sample. Next, I present simple correlations of response bias in the tasks without accounting for cognitive ability. Then, I show the results from the accuracy tests accounting for cognitive ability to confirm that a median split sufficiently divided the two ability levels; these results also provide a check of the manipulations. Lastly, I show the results from the response bias tests accounting for cognitive ability.

Measures of Cognitive Ability

Performance in the three cognitive ability tasks, Go/No-Go accuracy, Number-Letter shifting cost, and SAT/ACT test score percentile, are shown in Figure 4. The histograms show performance on the tasks, and the scatterplots show the relationships of the tasks with each other. The online sample, with the 179 included participants, is in the left column and the lab sample, with the 21 included participants, is in the right column.

Online/MTurk Participants	Lab Participants
	
	

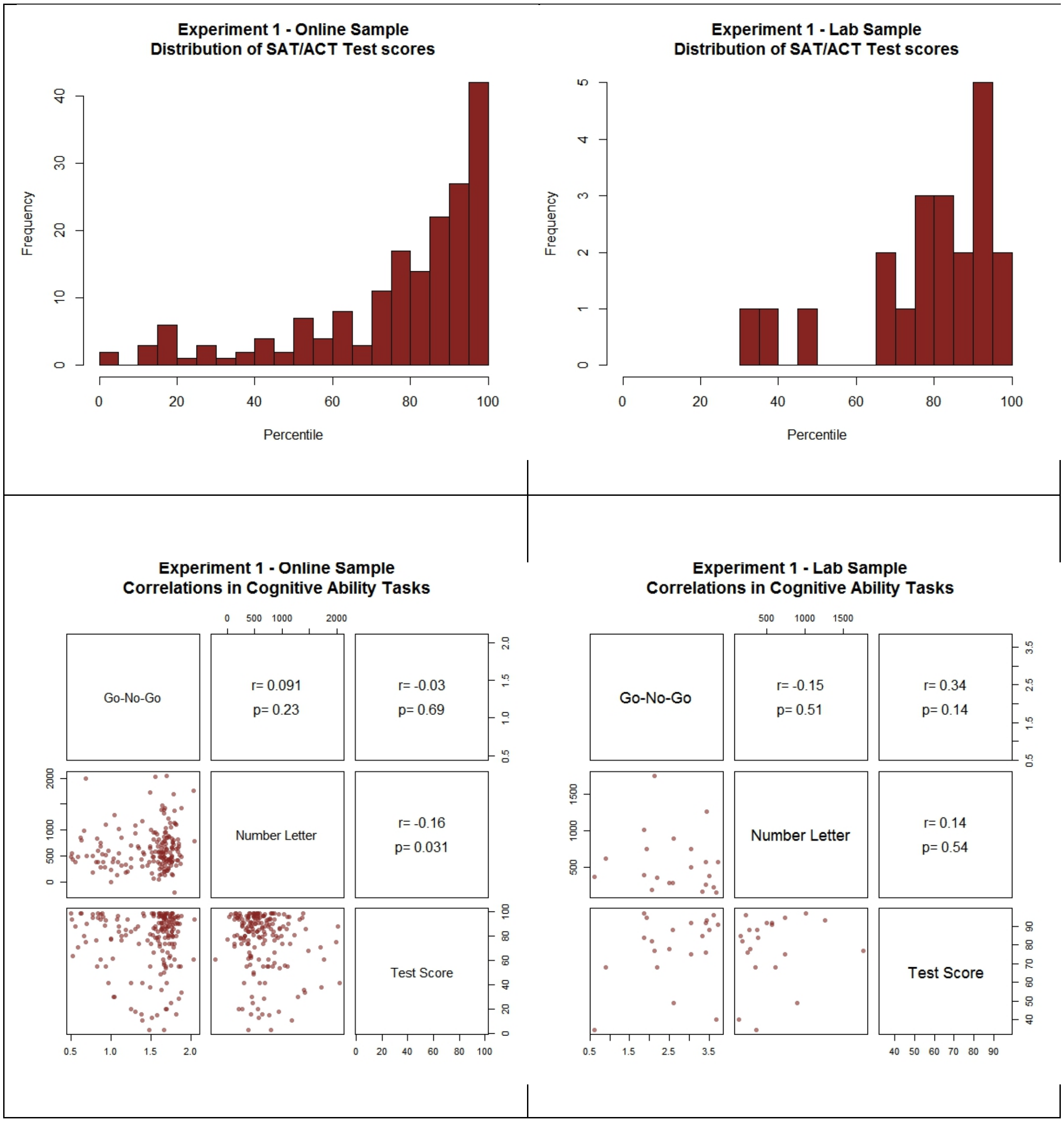


Figure 4. Histograms of performance on the cognitive ability task and scatterplots of their correlations in Experiment 1.

Go/No-Go task performance was not entirely consistent across the two samples. Notably, participants in the lab sample scored much better than those online; no online participant scored above *d'* = 2.5, but many participants in the lab had *d'* scores of 3-4. With perfect performance in this task yielding a *d'* of 4.48, such high scores in the lab sample may indicate ceiling effects, but it is also likely that the online participants lost motivation to perform well or attention in this task in a different way than lab participants who had the demand characteristics of participating in a lab with a research assistant in the next room. No feedback was gathered in Experiment 1, but in Experiment 2, which also included the Go/No-Go task, many online participants expressed their fatigue or frustration with the tedium in this task. Because this task was timed rather than self-paced, participants who grew fatigued or lost attention could simply stop responding and allow the task to finish. It seems unlikely, though, that the difference in performance on this task between the online and lab samples reflects a true difference in cognitive ability, based on the other metrics of cognitive ability included. And indeed, when participants did respond, those reaction times were very similar in the online sample and the lab sample (seen in Figure 5). While the discrepancy in the overall accuracy between the two samples is a concern for the validity of the measure, it appears to reflect a systematic difference in motivation rather than an actual difference in ability between the online and lab samples.

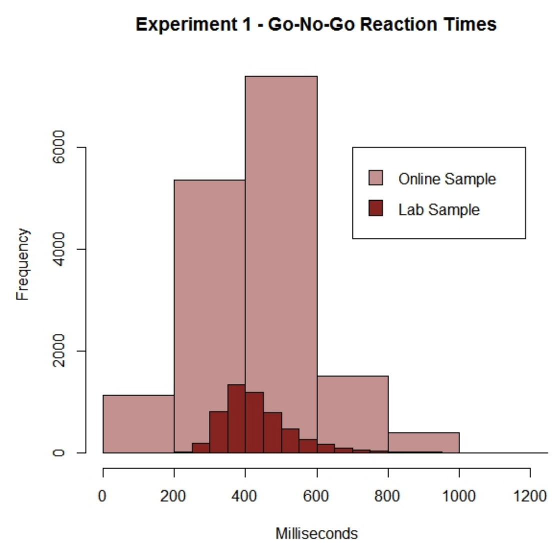


Figure 5. Distribution of reaction times in the Go/No-Go task in Experiment 1.

The distribution of SAT/ACT percentiles was also not the same in the two samples: there are many more scores in the lower tail in the online sample. Because admissions to universities usually require a minimum standardized test score, the distribution in the lab sample, taken from a university setting, *must* be truncated in the left tail in comparison to the online sample. Still, it was clear that online participants' reports of their scores were occasionally intentionally misrepresented, likely motivated by the fact that the SAT/ACT responses were a gatekeeper to whether participants could proceed to the rest of the task and therefore earn a much larger payment. The Qualtrics survey was intended to screen out impossible responses or unlikely years (say, if a 30 year old reported taking the SAT in 1994, when he would have been around 8 years old), and further screening criteria were implemented as mentioned above, but it is impossible to be certain that all individuals who provided inaccurate test scores were removed from the data.

More reassuringly, the discrepancies in the aforementioned cognitive ability tasks were *not* evident in the Number-Letter task. This may be an indication that the Number-Letter task had fewer artifacts, perhaps because the task was self-paced, and participants, regardless of sample, would likely be motivated to be finished with the study. Reaction times in the Number-Letter task were trimmed as Friedman et al. (2009) suggested: only correct responses with reaction time greater than 200 ms were included, and with further trimming for responses where the reaction time was more than 3.32 times more extreme than the median absolute deviation (i.e., 3.32 times the median value of the absolute deviations from the median) in the (shift or no shift) condition. This last criterion mostly trimmed out extremely long reaction times because median absolute deviation is a robust measure of outliers; in Gaussian distributions, the median absolute deviation is equivalent of 1.48 times the standard deviation (Rousseeuw & Croux, 1993), so using 3.32 times the median absolute deviation as a cut-off would remove values more extreme than roughly 4.91 standard deviations.

There was only one significant correlation among the three measures of cognitive ability in only the online sample: Number-Letter performance was significantly correlated with SAT/ACT score, $r(177) = -0.16$, $p = 0.03$ (seen in Figure 4). The lack of other significant correlations across the tasks could be further indication of the known measurement error in the SAT/ACT and Go/No-Go task in the online task, implicating issues with the current study, and could also be of concern because the tasks were used to measure cognitive ability together. However, the absence of clear evidence that these cognitive tasks measure the same underlying construct has also been reported in other studies. Miyake et al. (2000) did not use a conditional Go/No-Go task as was used in the current study, but they did include other measures of inhibition such as the anti-saccade (Roberts, Hager, & Heron, 1994), stop signal (Logan, 1994), and Stroop task (Stroop, 1935), and they found no significant correlations of those tasks with the Number-Letter task. Indeed the thrust of their paper was that the lack of significant correlations but significant structural equation model suggested that there were distinct aspects of executive functioning: inhibition was separate from shifting. It is very likely that the Go/No-Go task, which is a classic measure of response inhibition, maps onto the inhibition component of executive functioning; indeed, it is presumed to do so, for example, by the many studies that use the task to localize inhibition in imaging research (e.g., Simmonds, Pekar, & Mostofsky, 2008). While this does not eliminate the issues of validity in the current experiment, the general construct of cognitive ability as measured here is not necessarily problematic.

Measures of Response Bias

As a check on the effect of repetition and believability in the memory and reasoning tasks respectively, response bias was measured by the location of the criterion and the variability in the location of the criterion. In order to get these measures, the unequal-variance signal detection model was fit to the full ROC for each participant and each condition. While the memory data were easily fit to the unequal-variance signal detection model, this was not the case for the reasoning data. An explanation of the issue follows.

When participants do not use every confidence level, the ROC points are clustered. As explained in the introduction, each point is a cumulative sum of the hits and false alarms for all responses more conservative than that confidence level. For example, if a participant has a 0.25 hit rate and 0.12 false alarm rate using the most conservative confidence/”very sure old” level and 0.12 hit rate and 0 false alarm rate with the second most conservative confidence/”sure old” level, the first two points on the ROC, plotted false alarm rate on the x-axis and hit rate on the y-axis, would fall on (0.12, 0.25) and (0.12, 0.37). In other words, there would be 2 points in a vertical line. As this clustering increases due to unused confidence levels (for example, participants who only use the extremes of the scale), it becomes very difficult for the model to fit the data, including that the model is unable to appropriately determine the ratio of standard deviations from the target-to-lure distributions, i.e., the slope. Without this, the estimates of c_a are difficult to interpret.

Clustering of points on the ROC was a particular problem in the reasoning task. Because there were only 8 trials per item type in each condition, participants were unlikely to respond the full confidence scale (6-items) on each of these 8-trial cells, and indeed only about half of online participants used the full scale in the whole task at all, let alone for each condition and item type.

Because the estimates of c_a in the reasoning task were therefore flawed, three alternative approaches were considered. First, I could collapse across any confidence levels where that clustering would occur. This would give a different number of criteria than assuming all 6 confidence levels were used, but it would still allow me to analyze the relevant c_a as long as there were 3 confidence levels remaining (the minimum needed to estimate the slope). Unfortunately, over 93% of the reasoning data failed to meet this minimum requirement. Second, I could assume group participants into lower and higher cognitive ability and assume a common slope for the group. That is, I could collapse across the lower cognitive ability participants, find the slope of their ROC for the condition, and then use that slope to fit the individual data (and same for the higher cognitive ability participants). This would allow analyses of c_a, as in the memory task, but it would have the drawback of assuming that this slope is *constant* across participants in a given ability level. Third, I could assume an equal-variance signal detection model and analyze $c_.$ This would not assume differences were constant across *subgroups*, but it would have the drawback of assuming the slope is 1, the same for *all* participants. Much of the prior research using signal detection models for the belief bias task uses the unequal-variance model (e.g., Dube, Rotello, & Heit, 2010; Trippas, Handley, & Verde, 2013), but there is some recent research to indicate that it is appropriate to assume equal variance for this task (Trippas, Kellen, Pennycook, Fugelsang, & Koehler, 2016). This research is not yet peer-reviewed or published, though.

While none of these would be ideal solutions, the limitations in the data led me to assume the equal-variance signal detection model and use *c*. Specifically, when I attempted the group slopes approach, there was virtually no difference between using c_a and *c*: the group slopes only changed the calculations of c_a out to the thousandth decimal. Unfortunately, using group slopes did not resolve the main problem with these data: with participants using so few confidence levels, the ROC was nearly impossible to fit appropriately. The d_a estimates from the ROC fits would therefore not provide accurate estimates of discriminability. The only solution left was to assume the equal-variance signal detection model for the reasoning task, which is what will be presented.

Response bias was measured by absolute criterion location λ for all conditions and tasks and relative criterion location c_a/c for memory and reasoning respectively. For both λ and c_a/c, analyses were conducted on the location and on criterion location variability, measured with the range of criterion location calculated by subtracting the smallest from the largest. Because the memory task included two lists per condition, the standard deviation of criterion location across all four lists was also used as another measure of criterion location variability. In contrast, the reasoning task did not have multiple lists per condition, so a measure of standard deviation would provide the same information as the already included range. Accordingly, no analyses of the standard deviation in reasoning are presented.

Though not the major innovation of this experiment, there were significant correlations of relative criterion location, which replicated prior research. For example, previous findings show that response bias in memory is relatively stable across test lists despite up to week-long delays (Kantner & Lindsay, 2012, 2014). They measured relative response bias c_a, and indeed, the strong correlations of criterion location c_a in the four memory test lists (seen in Figure 6) clearly replicates this, though there was virtually no delay between lists in the current experiment. For both the online sample and the lab sample, there is a clear positive relationship in the scatterplots in the lower triangle of Figure 6, and the correlations of the criterion locations, seen in the upper triangle, are all highly significant.

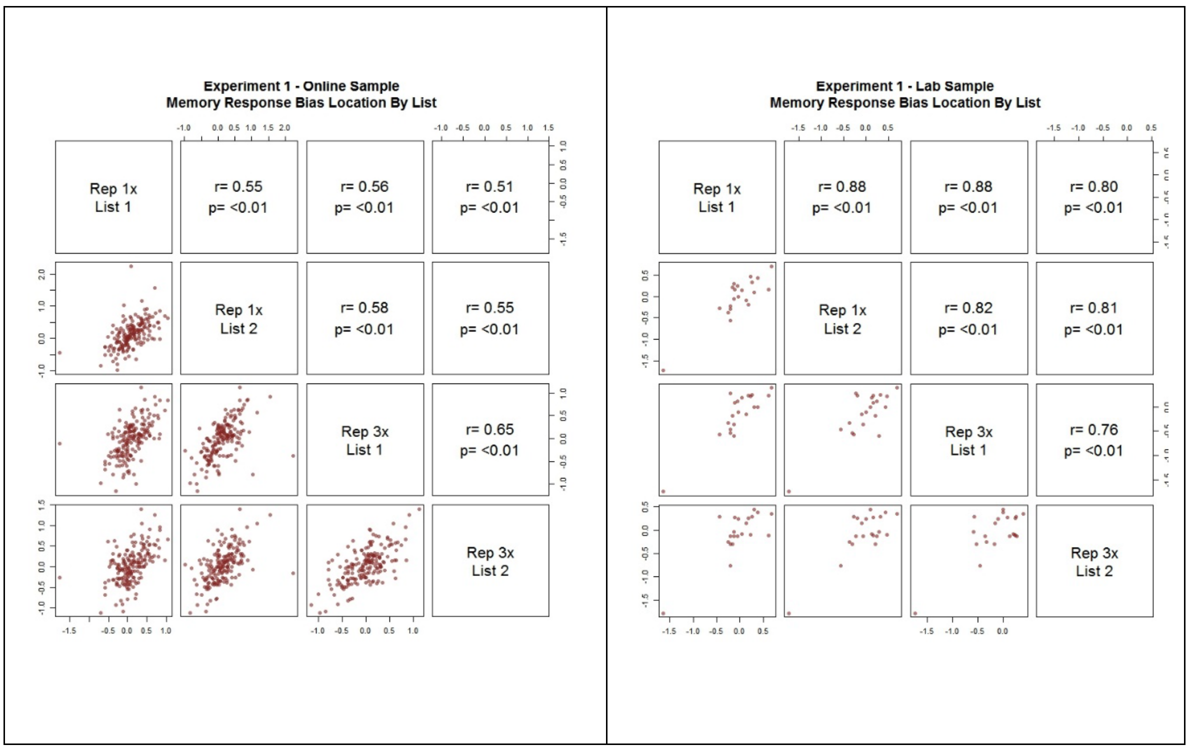


Figure 6. Scatterplot matrices of absolute response bias c_a in each of the four test lists in the memory task in Experiment 1.

Unlike in memory, relative response bias in reasoning was *not* correlated between the two conditions (see in Figure 7 the scatterplot in the third column/fourth row for the online sample and the lab sample respectively). This result is somewhat surprising given that the conditions were presented within the same list, suggesting that the relationship between the relative criterion location in the believable and unbelievable conditions may be more complicated than can be depicted with a simple correlation. Participants were expected to have a more liberal bias in the believable condition than the unbelievable (Dube et al., 2010); in the relevant scatterplots in Figure 7 as mentioned, this would mean points are mostly in the lower right hand portion of the plot, or points below a slope of 1 (i.e., a more liberal/lower numerical bias in the believable condition than the unbelievable condition). However, if the extent to which participants shift their bias is not consistent across participants, a correlation would *not* be expected to emerge. For example, a participant who has a very conservative bias in the unbelievable may shift a little bit (so their point would be in the upper right hand corner of the plot), or a lot (so their point would be in the lower right hand corner of the plot). Depending on the range of “starting” biases in the unbelievable condition and the size of shift, a simple correlation could show no relationship in overall location between the conditions. Indeed, this is further rationale to examine the individual differences that might explain response bias.

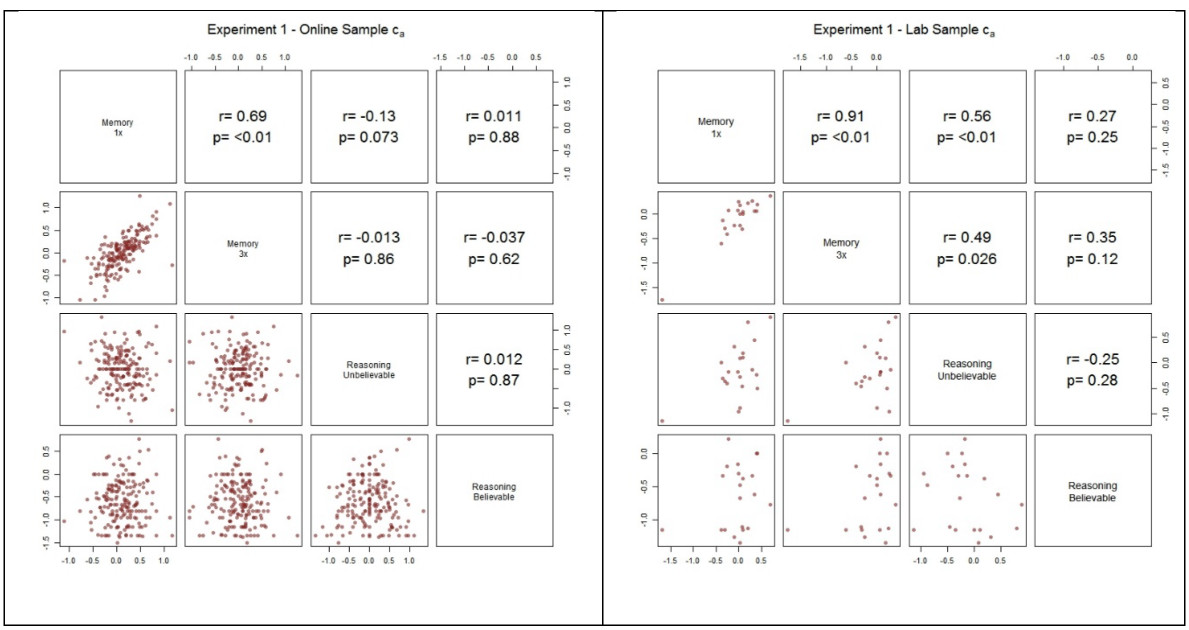


Figure 7. Scatterplot matrices of response bias in the memory and reasoning tasks in Experiment 1.

The correlations across memory and reasoning tasks care shown in the four scatterplots in the bottom left hand corners for the online and the lab samples in Figure 7. There were two significant correlations in only the lab sample: response bias in both the 1x memory condition and in the 3x memory condition correlated with unbelievable reasoning condition. However, there was an influential point that shows a very liberal criterion in both of these correlations (the point in the bottom left hand corner of each respective scatterplot). Without this outlying participant, the correlation is no longer significant (for the correlation of unbelievable reasoning response bias with the 1x memory response bias, $r(18) = 0.39$, $p = 0.08$, and for with the 3x memory response bias, $r(18) = 0.24$, $p = 0.31$).

Still, the correlations shown in Figure 7 did not account for cognitive ability as a potential predictor for individual differences in response bias: those analyses are presented next.

Accuracy Analyses

First, I checked that cognitive ability and condition led to the predicted effects on accuracy. That is, cognitive ability should lead to increased memory performance (e.g., Unsworth, 2010) and repetition should lead to higher memory accuracy, (e.g., Ebbinghaus, 1913), and higher cognitive ability participants should differ in their susceptibility to believable conclusions (Trippas et al., 2013; Trippas et al., 2014).

In order to test this, I used hierarchical linear models which included the within-subjects factor of condition as a binary predictor at the lowest level (level 1), and cognitive ability as a continuous predictor at level 2. β weights in the hierarchical model and interpretation of significant weights are presented below.

Memory

The measure of accuracy in the memory task was d_a. In the online sample, cognitive ability significantly predicted memory performance: higher cognitive ability participants had greater discriminability, $\beta = 0.21$, $t(177) = 2.49$, $p = 0.01$. However, this was not seen in the lab sample, $\beta = 0.19$, $t(19) = 1.24$, $p = 0.23$. This effect is likely underpowered given the small sample of only 21 participants, but it is in the same direction as the online sample. In both samples, there was also a significant effect of condition, as predicted, where more repetition led to greater discriminability (online sample: $\beta = 0.42$, $t(177) = 9.47$, $p < 0.001$, lab sample: $\beta = 0.60$, $t(19) = 6.26$, $p < 0.001$). Thus, condition predicted accuracy in both samples, but only the online sample had sufficient power to detect that cognitive ability predict accuracy.

Reasoning

In reasoning, the measure of accuracy was d'. In the online sample, higher cognitive ability participants had greater discriminability, $\beta = 0.34$, $t(177) = 3.10$, $p = 0.002$, and unbelievable conclusions led to higher discriminability than believable conclusions, $\beta = 0.32$, $t(177) = 4.43$, $p < 0.001$. The effect of believability on discriminability does run counter to Dube et al.'s (2010) findings that there was no effect on accuracy.

However, the key prediction here was that different cognitive ability participants would have different effects of believability on their accuracy (Trippas et al., 2013, 2014). This was not significant in the online sample: higher cognitive ability participants were only marginally significantly affected by the believability, $\beta = -0.20$, $t(177) = -1.66$,

$p < 0.10$. The results were different for the lab sample, though. Neither ability nor condition were significant alone, but there was a significant effect of ability on condition in the lab sample. Specifically, higher cognitive ability participants were significantly *more* affected by the believability of the conclusion, $\beta = 0.70$, $t(19) = 2.91$, $p = 0.009$. Indeed, this replicates Trippas and colleagues' (2013, 2014) accuracy findings that higher cognitive ability participants perform particularly better in the unbelievable condition, but no such benefit exists for lower cognitive ability participants.

Given the small size of the lab sample, it is surprising to see this differential effect of ability on the effect of believability accuracy with those participants, but not with the online participants. This could indicate that the online sample and lab sample here are not entirely comparable. While there is reason to believe that online samples are appropriate for cognitive psychology research (Paolacci, Chandler, & Ipeirotis, 2010), the discrepant findings in the current experiment could be a fault of the sampling technique or they could reflect meaningful differences in the populations. With the small lab sample, it is difficult to reach strong conclusions about the relationship between the lab and online data.

Response Bias Analyses

With indication that the measure of cognitive ability and the manipulations of repetition and believability were appropriate, I then analyzed the effect of cognitive ability on response bias. For both memory and reasoning, I conducted hierarchical linear models and simple regressions on the measures of response bias, depending on the number of relevant factors, respectively.

Memory

For response bias, measured with absolute criterion location λ, participants should be more conservative with increased repetition as Hirshman (1995) found; increasing repetition should improve discriminability enough that participants could minimize false alarms while still maintaining high accuracy. Additionally, I hypothesized that *higher* cognitive ability participants would be *more conservative* with *increased* repetition. However, this pattern was not observed for either absolute or relative measures of criterion location.

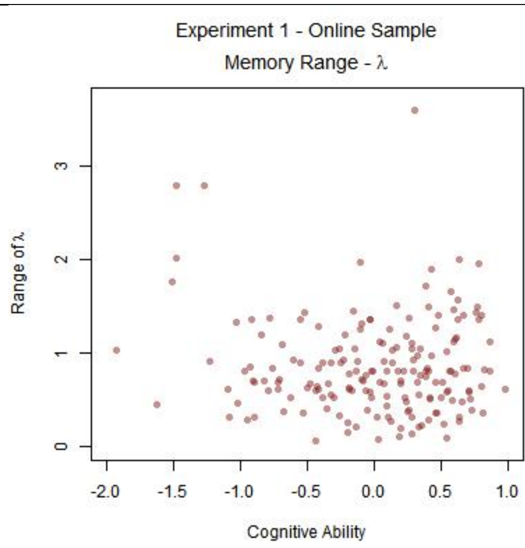
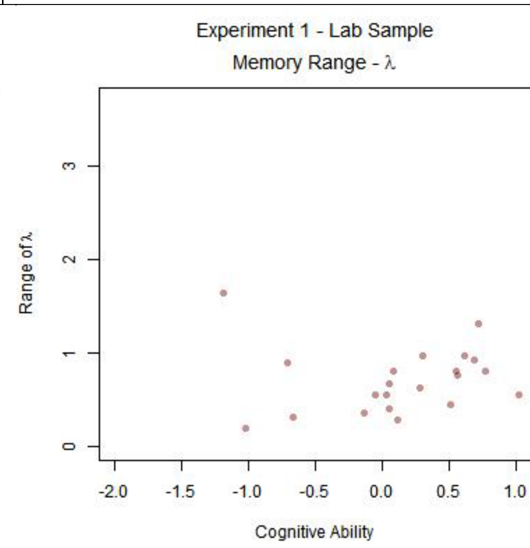
Using absolute criterion location λ, there was only a significant effect of condition: participants were more conservative with increased repetition (online sample: $\beta = 0.11$, $t(177) = 3.09$, $p = 0.002$, lab sample: $\beta = 0.27$, $t(19) = 4.36$, $p < 0.001$). No other effects were significant.

Using relative criterion location c_a, there was a significant effect of repetition, though in an unexpected direction: increased repetition led to a more *liberal* relative response bias in the online sample ($\beta = -0.10$, $t(177) = -4.78$, $p < 0.001$). This effect was marginally significant in the lab sample ($\beta = -0.07$, $t(19) = 1.79$, $p = 0.09$). No other effects were significant.

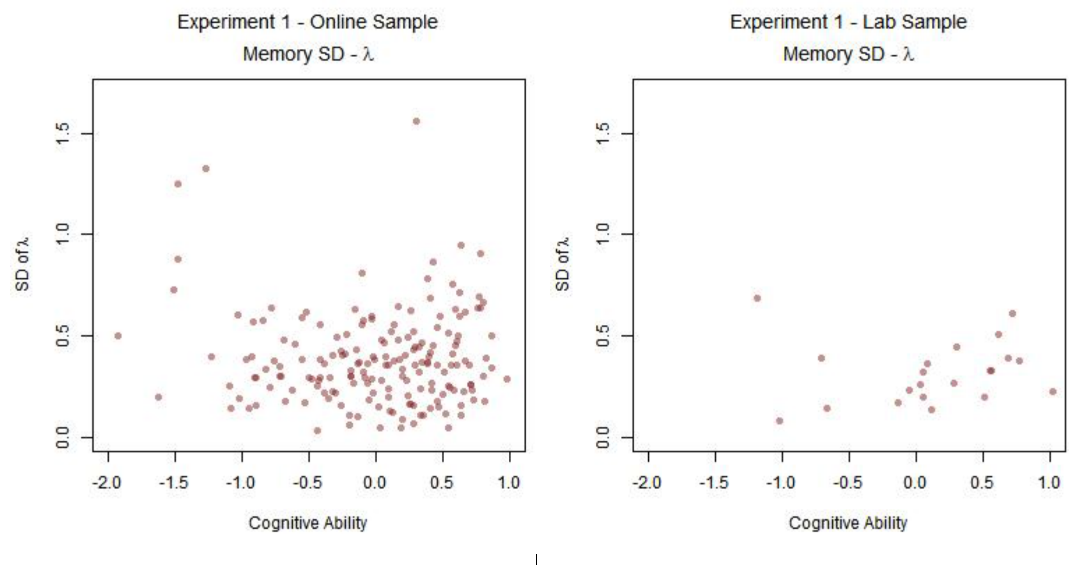
That the effect is in the opposite direction when using c_a at the measure of response bias in the online sample does not necessarily contradict the findings in λ. As was seen in the accuracy analyses, discriminability was not the same across conditions. When discriminability is not the same, the point at which $c_a = 0$ shifts its location along the evidence axis. Thus, even when λ becomes more conservative, to see a similarly conservative c_a, the shift towards the conservative decision bound must be greater than the baseline accounted for by the discriminability differences.

The predictions here were about absolute response bias, but relative response bias is important because the cognitive ability interactions predicted about absolute response bias were motivated by the suggestion that higher cognitive ability participants are better able to maximize optimality. A formal test of the optimality of response bias is further explored later in Experiment 2.

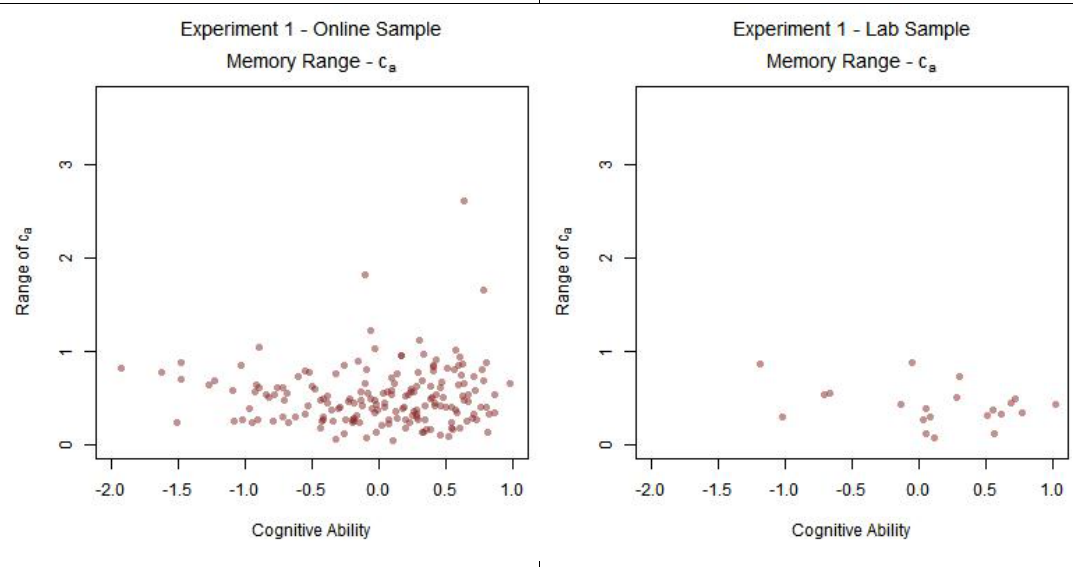
In addition to hypotheses of ability and criterion location, I also hypothesized that higher cognitive ability participants would shift their response bias more in order to appropriately adjust for the difficulty of the task. However, this was not the case for either measurement of range or standard deviation (sd) of either absolute criterion location λ or relative criterion c_a. This can be seen in the scatterplots in Figure 8 and the statistics in Table 1.

<table>
<tr><th></th><th>Online/MTurk Participants</th><th>Lab Participants</th></tr>
<tr><td>Range of λ</td><td>
</td><td>
</td></tr>
</table>

SD of λ



Range of c_a



SD of c_a

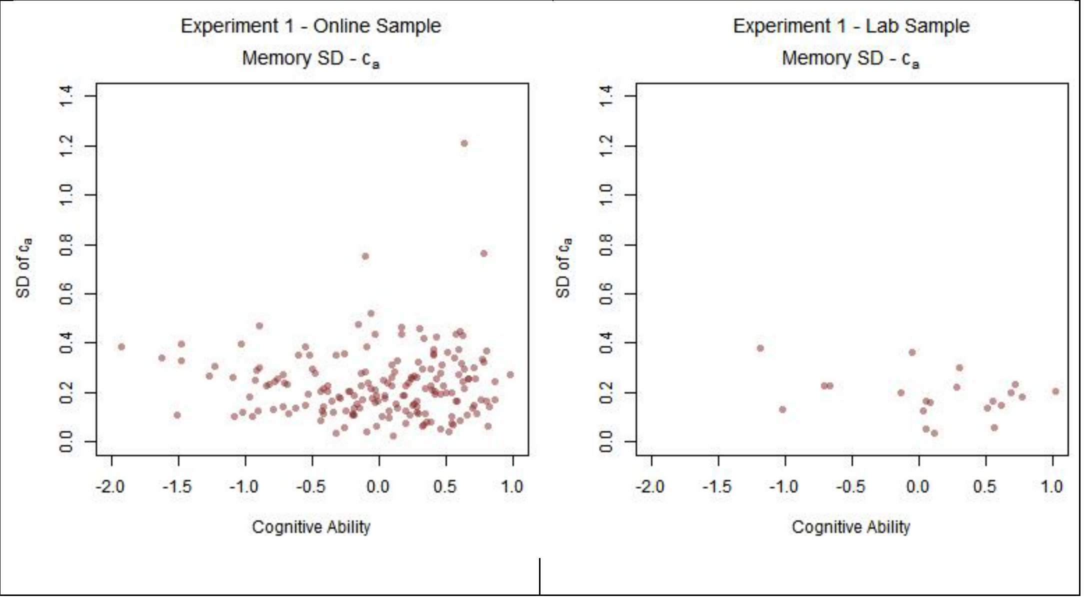


Figure 8. Plots of response bias variability, measured by range and standard deviation of λ and c_a at the old/new divide, in the memory task in Experiment 1.

			B	Significance
λ	range	Online Participants	-0.06	$t(177) = -0.85$, $p = 0.40$
		Lab Participants	0.03	$t(19) = 0.21$, $p = 0.84$
	sd	Online Participants	-0.02	$t(177) = -0.74$, $p = 0.46$
		Lab Participants	0.02	$t(19) = 0.41$, $p = 0.68$
c_a	range	Online Participants	0.02	$t(177) = 0.52$, $p = 0.61$
		Lab Participants	-0.11	$t(19) = -1.45$, $p = 0.16$
	sd	Online Participants	0.007	$t(177) = 0.40$, $p =0.69$
		Lab Participants	-0.04	$t(19) = -1.25$, $p = 0.23$

Table 1. Summary of simple regression of cognitive ability as a predictor of variability of response bias in memory in Experiment 1.

Reasoning

I hypothesized that there would be an interaction of cognitive ability and believability such that lower cognitive ability participants would be particularly more liberal in the believable condition. This was not the case.

Using absolute criterion location λ, in the online sample, participants were significantly more conservative in the unbelievable condition than the unbelievable condition in the online sample ($\beta = 0.87$, $t(177) = 13.03$, $p < 0.001$), and the lab sample ($\beta = 0.60$, $t(19) = 3.20$, $p = 0.005$), which replicated previous findings (Dube et al., 2010; Trippas et al., 2013). Higher cognitive ability participants were more conservative in the online sample ($\beta = 0.23$, $t(177) = 2.21$, $p = 0.03$), indicating a general effect of cognitive ability on response bias. This effect was marginal in the lab sample ($\beta = 0.50$, $t(19) = 1.83$, $p = 0.08$). Though not significant in the online same, there was also a marginally significant effect of ability on believability in the lab sample: higher cognitive ability participants were less affected by believability of the conclusion than lower cognitive ability participants ($\beta = 0.66$, $t(19) = 2.03$, $p = 0.06$). Thus, while I replicated the finding of more liberal response bias in the believable condition (Dube et al., 2010; Dube, Rotello, & Heit, 2011; Trippas et al., 2013), there was no clear evidence for an interaction as I hypothesized: the effect was only marginally significant and only in the lab sample.

Using relative criterion location *c*, unbelievable conclusions led to more conservative response bias than believable conclusions in both the online sample ($\beta = 0.71$, $t(177) = 12.18$, $p < 0.001$) and the lab sample ($\beta = 0.60$, $t(19) = 3.23$, $p = 0.004$). However, there were no other significant effects.

As with the memory task, I hypothesized that higher cognitive ability participants would shift their criterion in the reasoning task more to appropriately adjust for the difficulty of the task. However, this was not the case, as can be seen in the scatterplots in Figure 9 and in the statistics in Table 2. In the plot for range of λ in online participants, there are many points clustered in the lower right quadrant indicating participants with higher cognitive ability and smaller range. This was the only significant effect in shifting: higher cognitive ability participants shifted less.

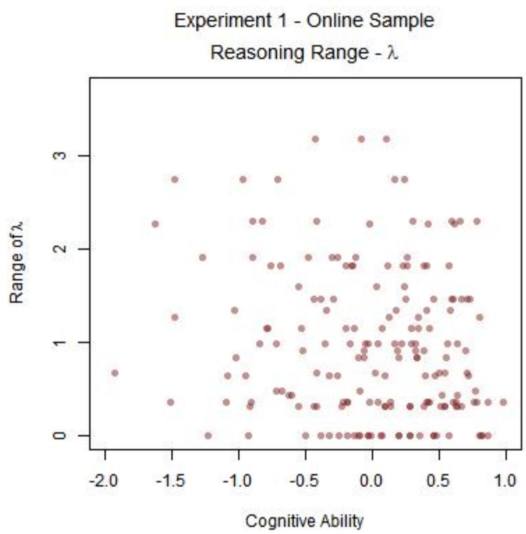
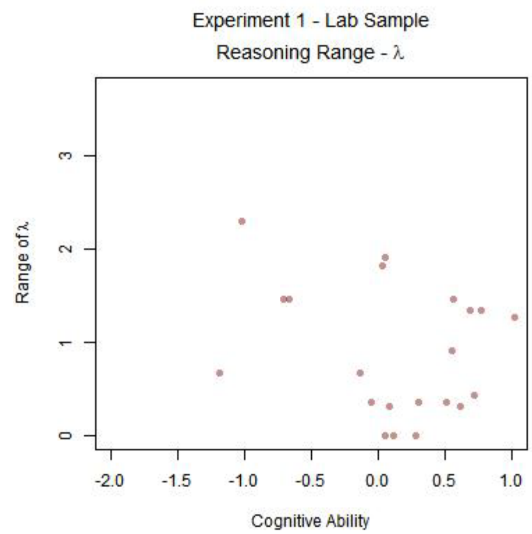
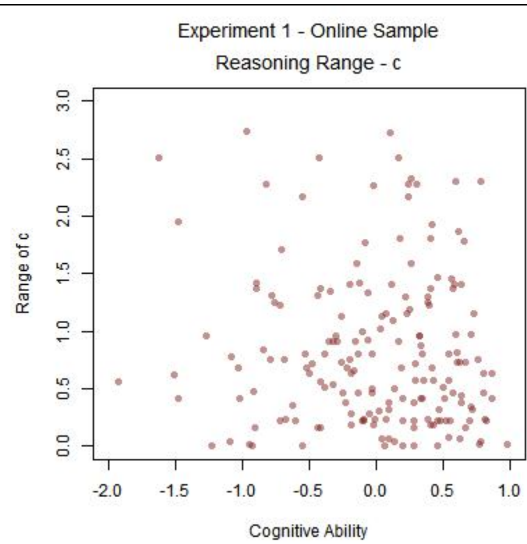
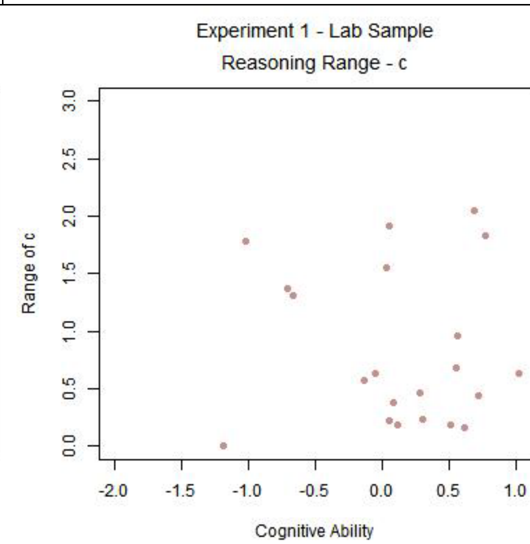
<table>
<tr><th></th><th>Online/MTurk Participants</th><th>Lab Participants</th></tr>
<tr><td>Range of λ</td><td>
</td><td>
</td></tr>
<tr><td>Range of c</td><td>
</td><td>
</td></tr>
</table>

Figure 9. Plots of response bias variability, measured by range of λ and range of c at the old/new divide, in the reasoning task in Experiment 1.

			B	Significance
λ	Range	Online Participants	-0.21	$t(177) = -2.04$, $p = 0.04$
		Lab Participants	-0.28	$t(19) = -1.08$, $p = 0.29$
c	Range	Online Participants	-0.09	$t(177) = -1.06$, $p = 0.29$
		Lab Participants	-0.09	$t(19) = -0.35$, $p = 0.73$

Table 2. Summary of simple regression of cognitive ability as a predictor of variability of response bias in reasoning in Experiment 1.

Experiment 1 Discussion

While there were indications of some differences between the online and lab samples (different accuracy performance on the Go/No-Go task and greater spread in the SAT/ACT scores), these likely indicate differences in motivation rather than in cognitive ability. Other measures did appear to be qualititatively similar (similar distribution of reaction times in the Number-Letter task and the Go/No-Go task), so it was deemed appropriate to use both samples.

Examining effects on accuracy provided a manipulation check and indicated, as predicted, that higher cognitive ability participants performed better in both the memory and reasoning task, and repetition led to greater discriminability in memory, and believable conclusions led to lower discriminability in reasoning. There was an interaction of ability and believability in the lab sample, consistent with conclusions from prior research (Trippas, Verde, et al., 2014), but not in the online sample. This could reflect that cognitive ability is a very broad term used to describe multiple constructs because they used different kinds of tasks than the ones used here. This could also potentially reflect differences in the online and lab sample. However, the effect of ability on accuracy indicates that it is a relevant individual difference here.

The main hypotheses in the current experiment were on response bias. Without accounting for cognitive ability, response bias remained stable in memory across lists, replicating prior research (Kantner & Lindsay, 2012). Bias shifts due to the manipulation in the tasks were significant and as expected: participants were more conservative in their absolute response bias in the stronger memory strength condition, replicating Hirshman (1995), and were more liberal in the believable condition, replicating Dube et al. (2010). A stronger test of bias shifting, using relative response bias, did not result in the same effects in memory, but did in reasoning. Additionally, I predicted an interaction of cognitive ability and condition that did not emerge. Indeed, there is no evidence to indicate cognitive ability played a role in response bias at all in memory. However, in reasoning, higher cognitive ability participants were more conservative and shifted their bias less than lower ability participants.

While both absolute and relative response bias measures were presented, they do have different implications, as mentioned earlier. In the EVSD case, c is relative to the midpoint between the distributions, which is also the optimal criterion that would maximize accuracy. In the UVSD case, c_a is relative to the midpoint between the distributions, but that is *not* the same as optimal criterion because of the difference in variances. Looking at response bias in comparison to optimal bias for individuals would be particularly interesting. If optimal response bias is different in different conditions, it is possible that participants responding optimally would shift their bias more than participants not responding optimally. A closer examination of the extent to which participants shift their bias and how cognitive ability may explain that is presented next in Experiment 2.

CHAPTER 3

EXPERIMENT 2

Given the consistent evidence in previous research and also in Experiment 1 that cognitive ability appears to influence response bias, it was further studied in conjunction with bias manipulations, since determining the appropriate response bias may differ depending on the situation and the individual. Experiment 2 expanded on the study by Aminoff and colleagues (2012) looking at individual differences and bias manipulations. While several cognitive tasks were included in Aminoff and colleagues' study, they measured mostly verbal cognitive ability, rather than executive functioning aspects of cognitive ability.

Experiment 2 focused on exploring how participants might choose to respond optimally if there are competing definitions of optimality: maximizing accuracy, or maximizing the expected value of the items. In other words, optimal criterion could be to maximize correct responses and minimize errors when hits and correct rejections are equally weighted (yielding a likelihood ratio of close to 1). But optimal criterion could be to maximize correct responses and minimize errors when the benefits of hits and correct rejections are *not* equally weighted (yielding a likelihood ratio less than 1 and more liberal response bias if hits are particularly beneficial, and a likelihood ratio greater than 1 and more conservative response bias if correct rejections are particularly beneficial); this would maximize the expected value. In particular, when the weights are externally determined, say via a payoff manipulation in which different point values are given for hits and correct rejections, an optimal criterion would be one that maximizes the number of earned points. In such a condition, one cannot both maximize accuracy *and* points with the same decision rule. When there is a tension between these optimal decision rules with a payoff manipulation, participants will compromise and choose a decision rule between the two (Bohil & Maddox, 2001; Maddox & Bohil, 2004). Understanding how individual differences in cognitive ability interact with payoff manipulation was the main goal of Experiment 2. Particularly, how do participants determine decision criteria when there is competition in optimal decision rules, and are higher cognitive ability participants better able to determine optimal decision criteria when instructed ahead of time which decision rules to maximize?

Method

Participants

Experiment 2 was conducted entirely online for feasibility, given the sample size needed for sufficient power. Again, participants were recruited through Amazon.com's Mechanical Turk and were restricted to ages 18-30, with the same requirements and process to getting to the "bonus" study as in Experiment 1, except with the additional requirement of not having participated in the first experiment. Also, participants were told they could earn up to an additional 72 cents for their performance on the task in addition to earning the $3 for completing the "bonus" study. This was to incentivize the payoff manipulation. Of the 898 participants who completed a prescreening to determine eligibility, 688 participants were eligible and 369 completed the full experiment.

Participants were excluded from analysis based on the same criteria as in Experiment 1. There were 18 participants who reported an implausible SAT/ACT score but were still able to access and complete the experiment. An additional 30 participants performed poorly on the Go/No-Go task (lower than d' of 0.5). As in Experiment 1, no participants were excluded based in their median reaction time in the Number-Letter task (with cut-off of either less than 800 ms or greater than 2000 ms). There were 59 participants whose accuracy was not at least 10% greater for their highest confidence responses than their lowest confidence responses. With these participants excluded, a remaining 262 participants were included in the full analyses.

Because of the many manipulations that were included for the memory task and to ensure a reasonable study duration, the instructions manipulation (Maximize Accuracy, Maximize Payoff, and Maximize Both) was between-subjects. Specifically, in the final sample presented below, 90 participants were in the Maximize Accuracy Instructions condition, 88 were in the Maximize Payoff Instructions condition, and 84 were in the Maximize Both Instructions condition. Further explanation of these conditions will appear in the Memory task section.

Cognitive Ability Measures. As in Experiment 1, I measured cognitive ability with self-reported SAT or ACT, the Go/No-Go task, and the Number-Letter task. Because there was some concern about measurement error of these tasks in Experiment 1, I also added the Cognitive Reflection Test (Frederick, 2005) with some modification. The Cognitive Reflection Test is a three question measure of cognitive ability that Frederick used to predict decision making. Because some participants may have encountered the task previously (for example, in another online study on Mechanical Turk) and may have memorized the answers, I changed some of the wording. For example, one original test question reads, “If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?” In my version, I asked how long it would take 100 machines to make 50 widgets. If a participant did not know the correct answer to the original problem, they might respond 100 or 50 minutes; if a participant had memorized the correct answer to the original problem and did not read carefully, they would respond 5 minutes. The correct answer here is 2.5 minutes.

Memory Task

Instead of including two tasks like in Experiment 1, only the memory task was implemented in Experiment 2, given the complexity of the design and the difficulty in having a sufficient number of reasoning problems. This memory task included three study lists each of 40 words, each word presented three times. The test lists each included 40 targets and 40 lures.

In order to examine the relationship between optimal responding and bias manipulations, I used a payoff manipulation including conservative, neutral, and liberal payoffs. The payoff manipulation was implemented between lists and, like in some other studies (e.g., Bohil & Maddox, 2001), consisted of only rewards for correct responses (no penalties for incorrect). This was done in part to minimize confusion in the online sample because they would not be able to receive further explanation beyond the written instructions.

For the liberal payoff condition, participants saw Figure 10 on screen for the duration of that list. This figure shows that hits would be rewarded more points than correct rejections. In other words, they should try to be liberal in responding that an item is old in order to earn the most number of points.

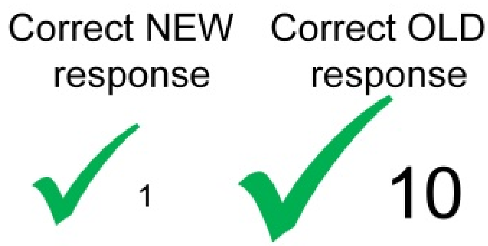


Figure 10. A liberal payoff manipulation. In this case, responding “old” correctly gives a larger reward. Responding “new” correctly gives a smaller reward. If participants are maximizing the reward, they should be biased to respond “old” to the item.

There was also a conservative payoff condition which rewarded correct rejections more than hits and a neutral payoff condition in which participants earned the same amount regardless of type of correct response. Order of the manipulation was counterbalanced.

One third of the participants saw instructions about the payoff, and that accuracy and the payoff are both important (Maximize Both), as is commonly done in payoff manipulations. Because the base rates remained 50% old and 50% new items at test as in Experiment 1, participants would not be able to choose a decision rule that would maximize both accuracy and expected value, but they could choose to compromise (Bohil & Maddox, 2001). In order to better understand how participants choose their criteria, the other participants were instructed towards a decision rule. One third of the participants saw instructions that asked them to Maximize Accuracy because the amount of extra money they could earn would be based on their accuracy. The last third of the participants saw instructions that asked them to Maximize Points because the amount of extra money they could earn would be based on the number of points at the end. Instead of using cash in the payoff, points were converted to a cash equivalent proportional to the possible 72 cent performance-based bonus.

By having different payoffs, participants attempting to find an optimal response criterion need to weigh accuracy and payoff in order to do so and because they may choose to sacrifice some accuracy for payoffs, may ultimately decide on a compromise (Bohil & Maddox, 2001). By including different instructions emphasizing accuracy vs. points/rewards, I examined how participants are influenced by this tension, by goals from the instructions, and whether higher cognitive participants are better at choosing the appropriate response bias.

Results

Variable Measurement, Approach to Analyses, and Hypotheses

There were two main dependent variables of interest: one, the individual's absolute response criterion location λ and two, how optimal the individual's criterion location was (the difference between the individual's absolute response bias location λ and the optimal response bias location). To define optimal response bias when the benefits of hits and correct rejections are not the same (the conservative and liberal payoffs here), the likelihood ratio formula (in Equation 3) was used. However, to provide easier comparison to Experiment 1, the optimal likelihood ratios were converted to an absolute location on the x-axis (i.e., transformed to the same scale as λ) and then subtracted from the individual's response criterion location λ. This means that interpretations of the numbers are of criterion location on the x-axis (which were presented in Experiment 1), rather than of likelihood ratios (which were not).

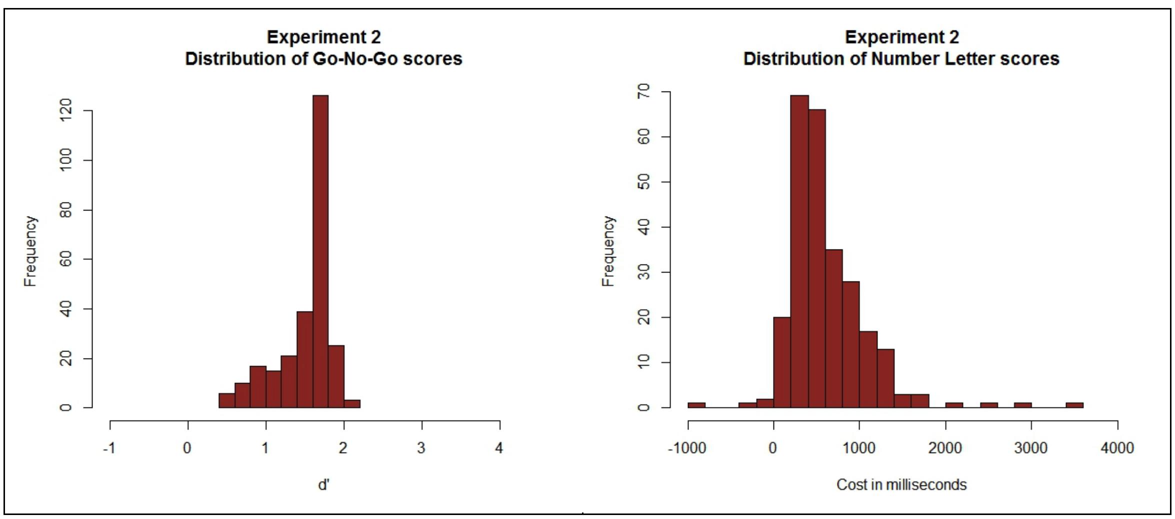
<u>Approach to Analyses, and Hypotheses</u>

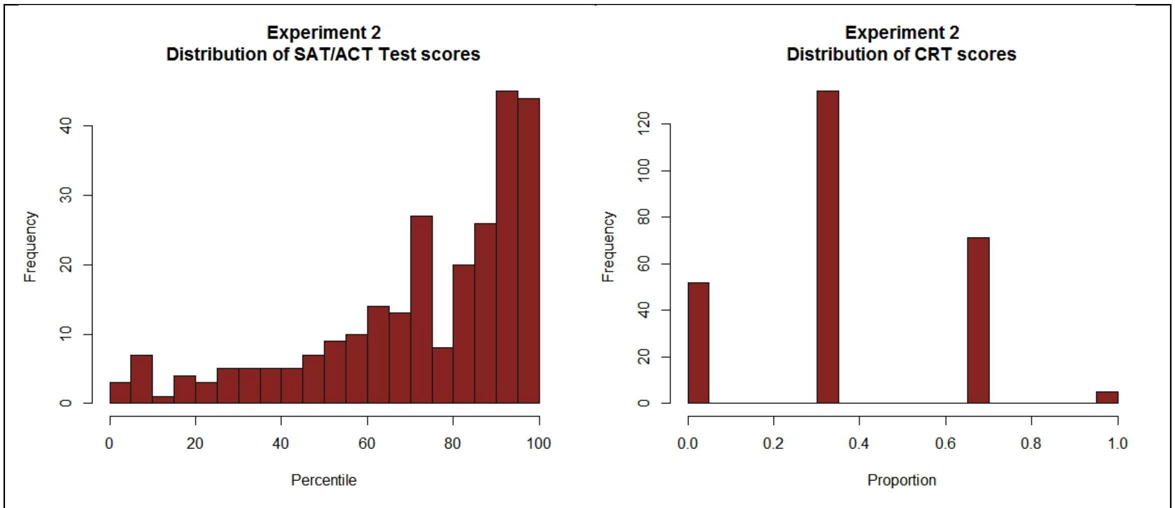
Participants were grouped into higher and lower cognitive using a median split from a composite score of the component ability tasks (see Trippas, Handley, & Verde, 2013, 2014 for precedence for using a median split). A 2x3x3 mixed ANOVA (higher/ lower cognitive ability and payoff/accuracy/both instruction manipulations between subjects, conservative/neutral/liberal payoff manipulations within subjects) was conducted on the dependent variables of interest. I hypothesized that participants would be influenced by the instructions: when told to Maximize Points, participants would generally try to Maximize Points. I also hypothesized *that higher cognitive ability participants* would be *closer to optimal* than lower cognitive ability participants. So when the instructions were told to Maximize Points, higher cognitive ability participants would set a criterion to optimize their points and when the instructions were to Maximize Accuracy, higher cognitive ability participants would set a criterion to maximize their accuracy. I also predicted that the tension between optimal according to the payoff and optimal according to accuracy should be evident in the conservative and liberal conditions (where the two conflict) and lead to suboptimal response bias, but not in the neutral condition (where they coincide). I also hypothesized that this tension in combination with different goals from the instructions would lead to participants trying to resolve towards the goals from the instructions, but that it would be easier for higher cognitive ability participants. That is, *higher cognitive ability participants would have closer to optimal response bias that maximizes their instructed goal when there is conflict* in how optimal could be defined. For example, I predicted that when told to Maximize Points, higher cognitive ability participants would have a closer to optimal response bias for points in the conservative and liberal payoffs. An interaction of cognitive ability and payoff manipulation would indicate that future experiments that use a payoff manipulation should include cognitive ability, perhaps as a covariate, to account for what has historically been considered acceptable noise.

The results from the three measures of cognitive ability are presented first. Next, I present the ROCs and the results from the accuracy tests accounting for cognitive ability to confirm that a median split sufficiently divided the two ability levels. Then, I present the results from the response bias tests accounting for cognitive ability first with absolute response bias location and then with the deviation from optimal response bias location.

Measures of Cognitive Ability

Performance in the four cognitive ability tasks, Go/No-Go accuracy, Number-Letter shifting cost, and SAT/ACT test score percentile, and proportion correct on the CRT are shown in Figure 11. The histograms show performance on the tasks, and the scatterplots show the relationships of the tasks with each other.




Experiment 2
Distribution of SAT/ACT Test scores
Frequency
Percentile
0
10
20
30
40
60
80
100
Experiment 2
Distribution of CRT scores
Frequency
Proportion
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0

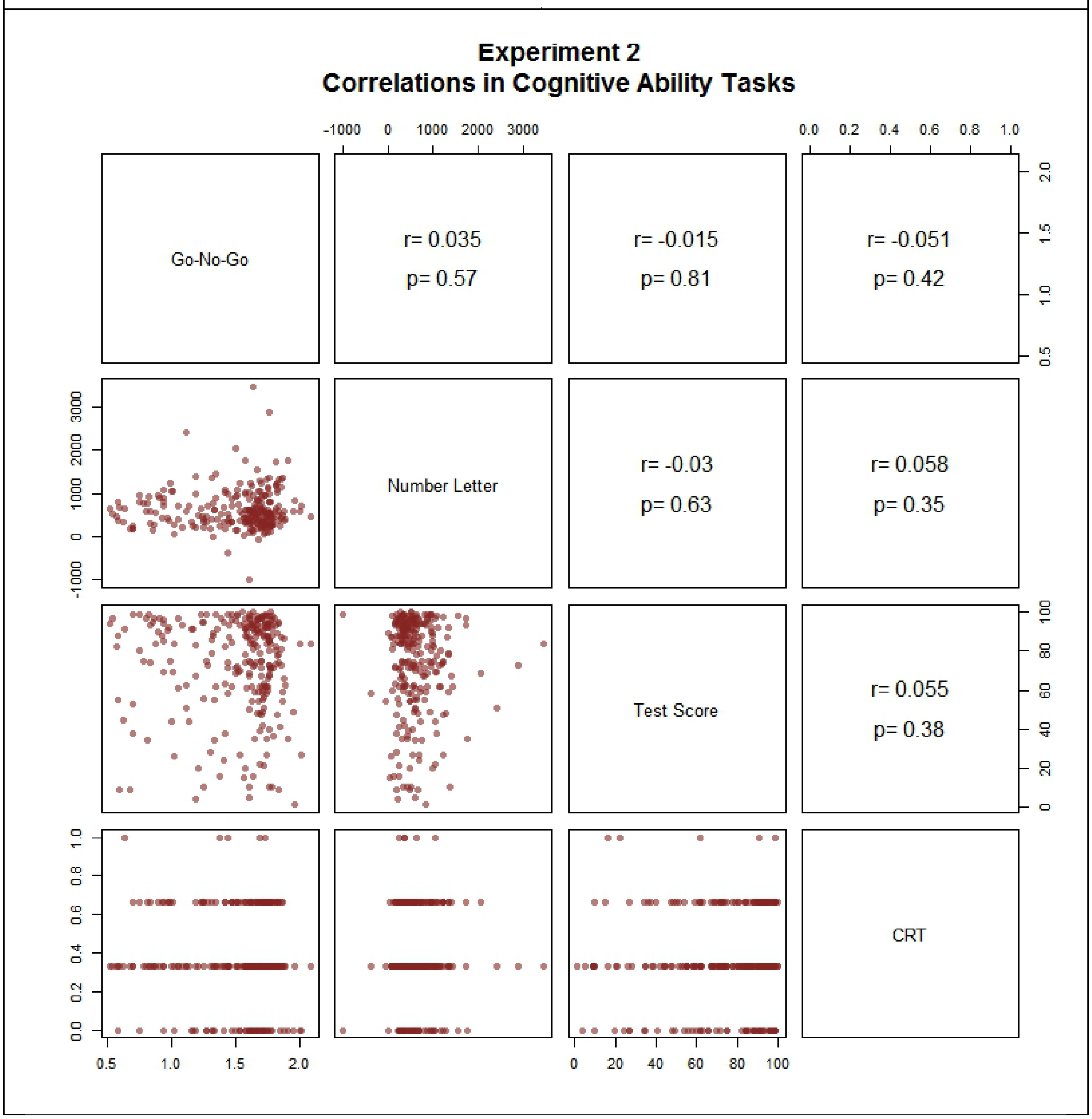

Experiment 2
Correlations in Cognitive Ability Tasks
Go-No-Go
r= 0.035
p= 0.57
r= -0.015
p= 0.81
r= -0.051
p= 0.42
Number Letter
r= -0.03
p= 0.63
r= 0.058
p= 0.35
Test Score
r= 0.055
p= 0.38
CRT
-1000
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
20
40
60
80
100

Figure 11. Histograms of performance on the cognitive ability tasks and scatterplots of their correlations in Experiment 2.

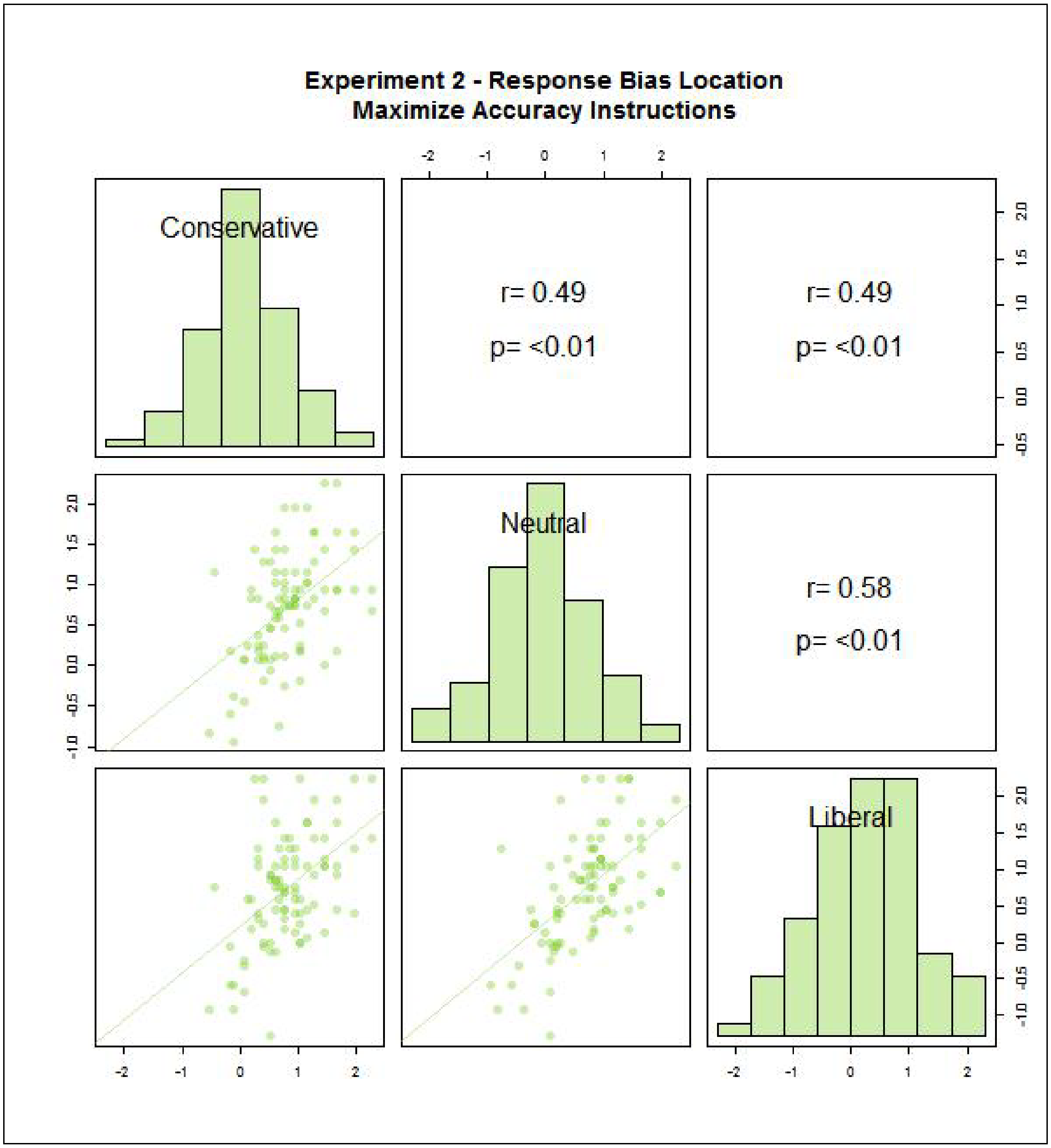
Visually comparing these scores (Figure 11) to those in the online sample from Experiment 1 (Figure 4) reveals a high degree of similarity. Go/No-Go task performance had a similar upper "limit", and SAT/ACT performance included a left skewed distribution though with non-negligible amounts of very low scores. The histogram for Number-Letter task performance had a similar mode and though the tails were a little longer in Experiment 2 with about 4 participants responding with much longer mean RT than in Experiment 1.

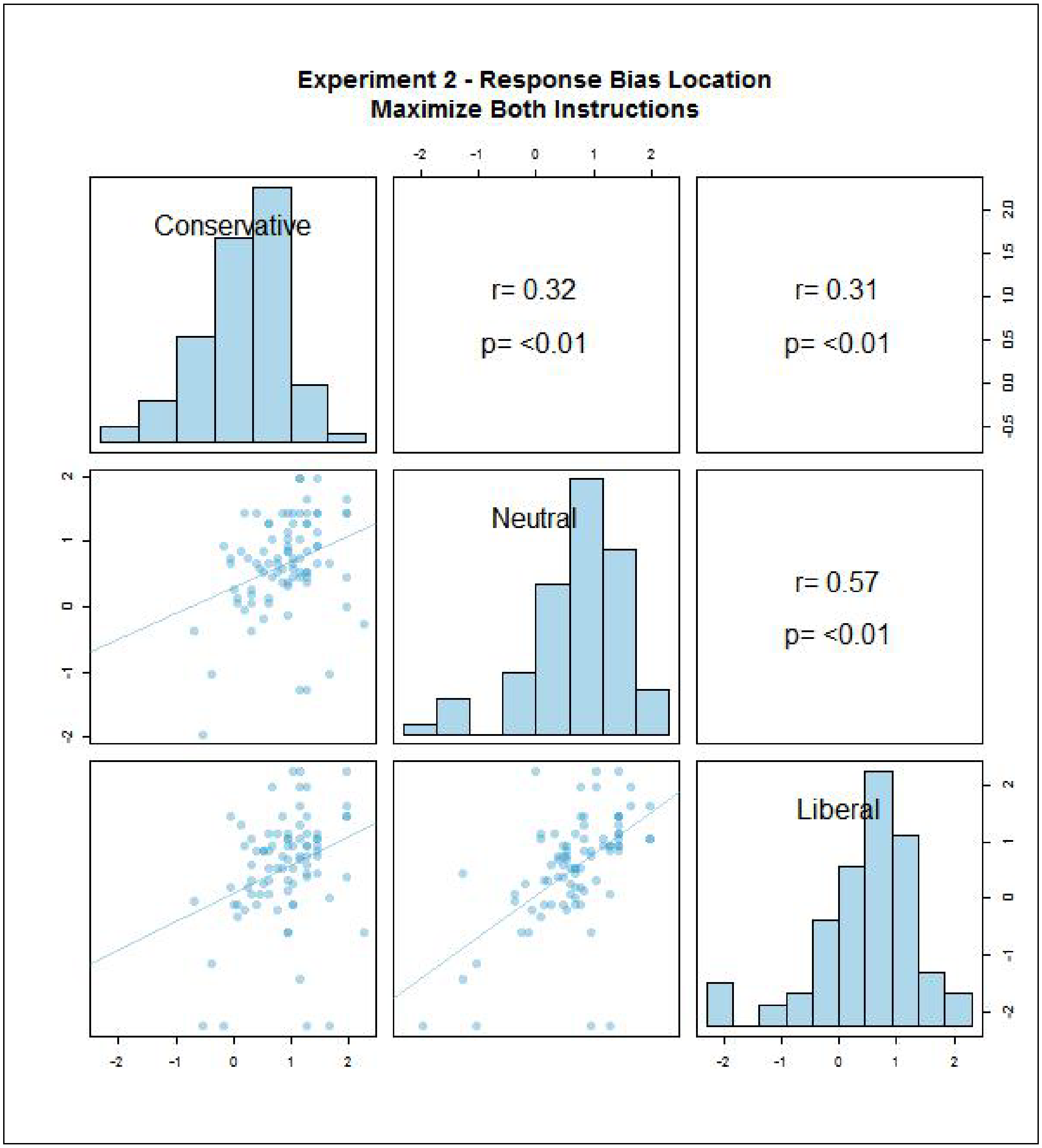
The Cognitive Reflection Test, used only in Experiment 2, showed a right skew in performance: most participants either missed all three items or only got one correct. Including the CRT as a measure of cognitive ability for a composite-score based median split changed less than 18% of participants' status from higher to lower cognitive ability or vice versa. It is possible that the inclusion of the CRT reduced the impact of any misleading self-reported SAT/ACT scores as was the concern in Experiment 2, if only by simply weighting each task less in the composite score-based median split.

As with Experiment 1, there were no significant correlations among the three measures of cognitive ability. There were there no significant correlations with CRT as well. In short, this sample was very similar to the online sample in Experiment 1, as would be expected.

Measures of Response Bias and Response Bias Optimality

Absolute response criteria across different payoff conditions in Experiment 2 were highly correlated for participants in each instruction condition (seen in Figure 12). The one exception was that participants in the Maximize Points instructions condition did not have significantly correlated response bias locations in the conservative and liberal payoff conditions, $r(86) = 0.16$, p = 0.13. Shifts in bias are difficult to see in the histograms here, but they are clearer in the ANOVA presented later and do indicate that participants were appropriately influenced by the payoff conditions. These correlations shown in Figure 12 did not account for cognitive ability as a potential predictor for individual differences in response bias: those analyses are presented next.




Experiment 2 - Response Bias Location
Maximize Both Instructions
Conservative
r= 0.32
p= <0.01
r= 0.31
p= <0.01
Neutral
r= 0.57
p= <0.01
Liberal

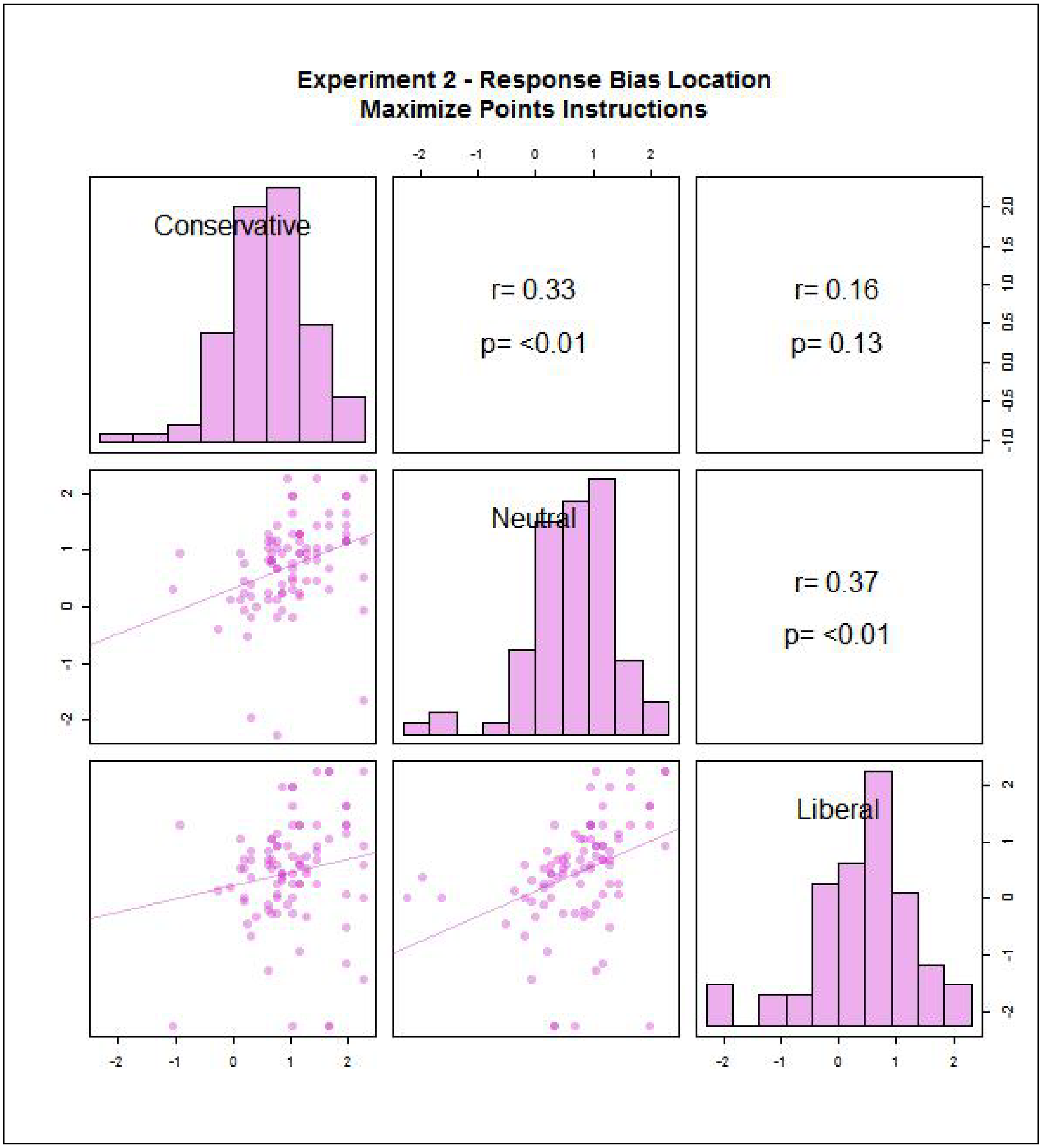
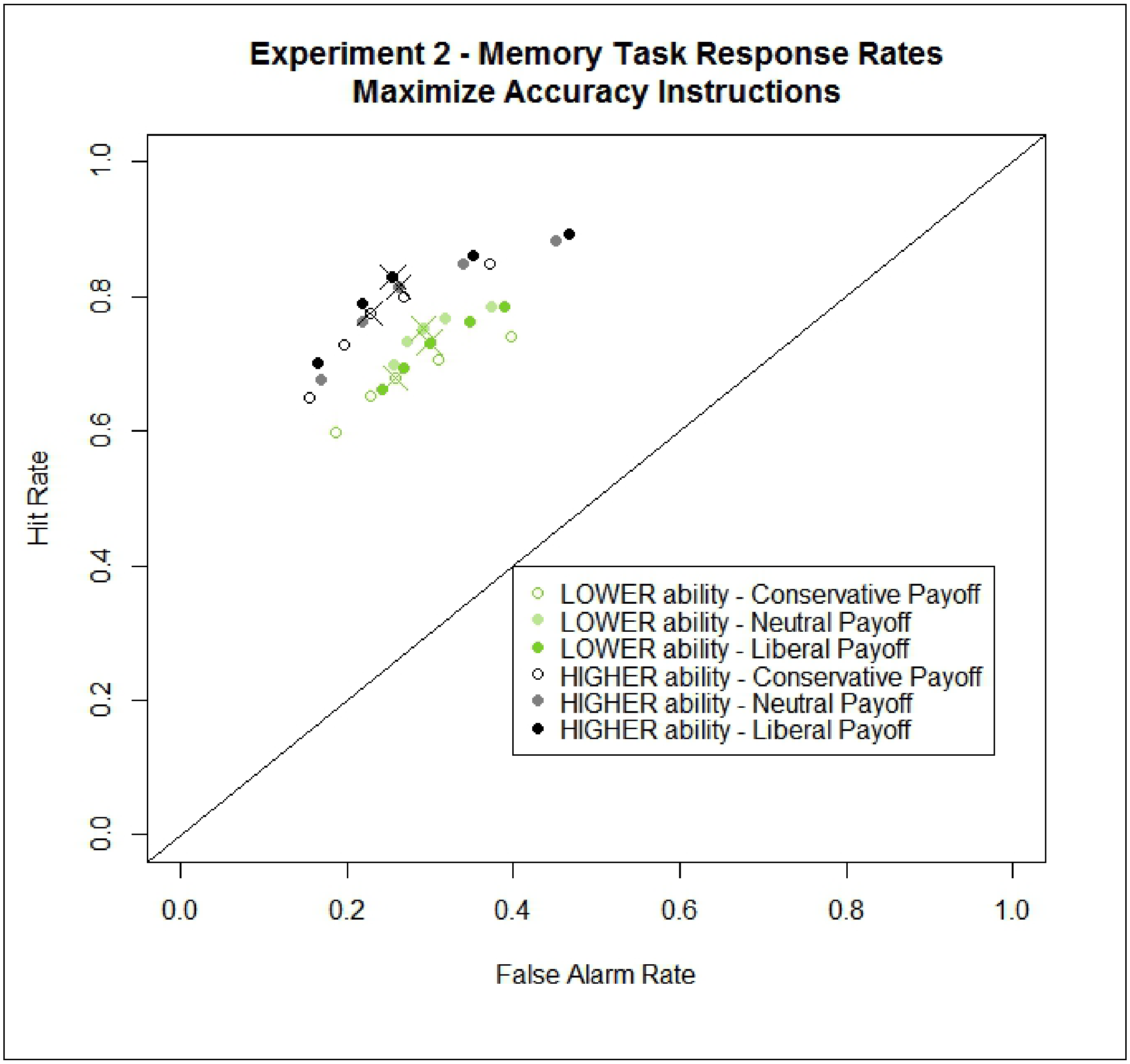


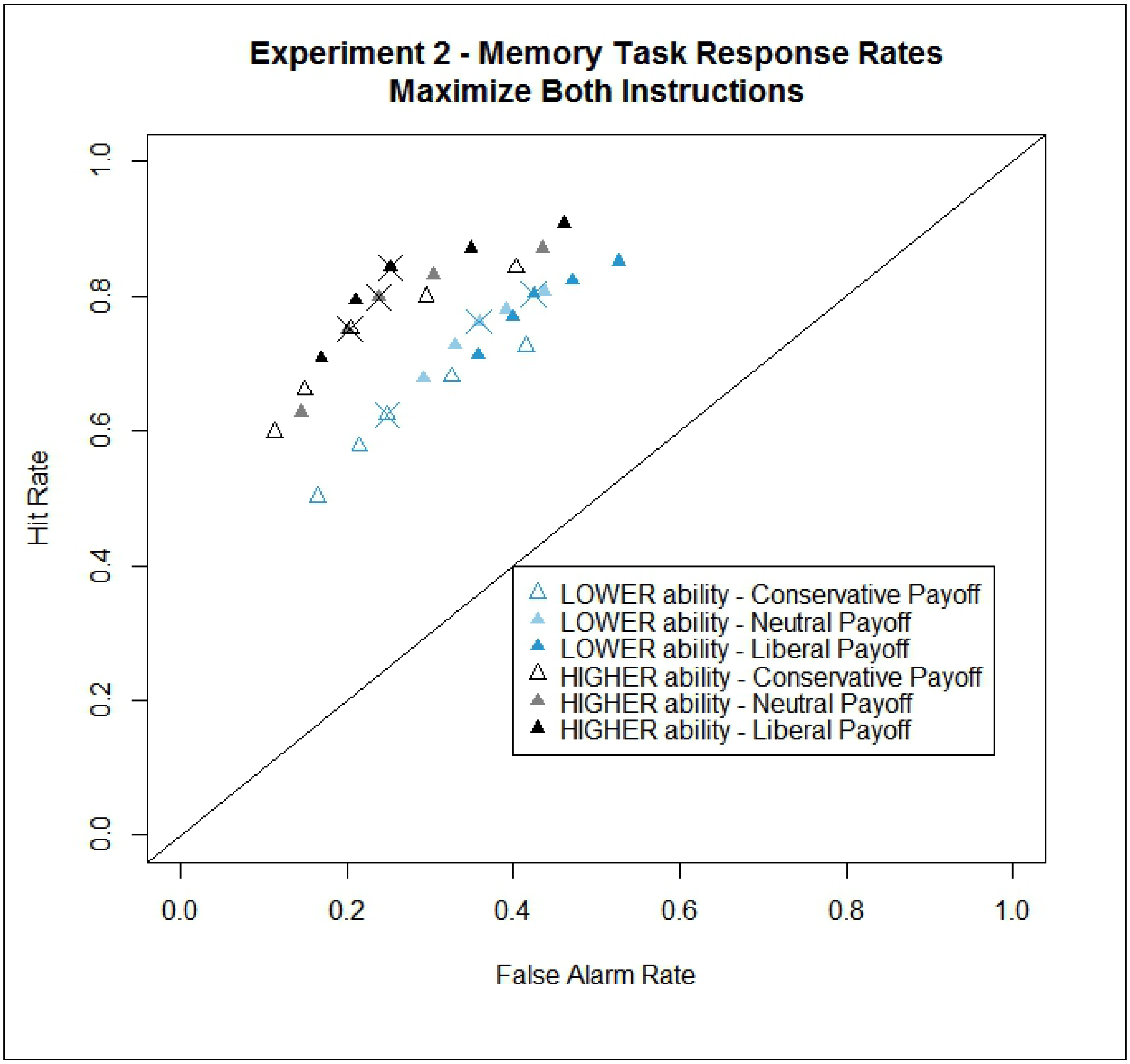
Figure 12. Scatterplot matrices of response bias in the memory and reasoning tasks in Experiment 2.

Median Split, ROCs, and Accuracy Analyses

Median Split and ROCs

The ROCs for the memory task are shown in Figure 13 in order to provide an overview of the data. There are three separate graphs for each of the instructions conditions, which were manipulated between subjects. Each graph includes the within-subjects manipulation of payoff, where the darker the symbol, the more liberal the payoff. Higher cognitive ability participants are shown in black in the ROC, with lower cognitive ability participants in green for the Maximize Accuracy instructions, blue for the Maximize Both instructions, and purple for the Maximize Points instructions. Again, points with an x through them depict the criterion location that divides “old” responses from “new” responses; these points were used for the response bias analyses presented later.





Experiment 2 - Memory Task Response Rates
Maximize Both Instructions
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0
LOWER ability - Conservative Payoff
LOWER ability - Neutral Payoff
LOWER ability - Liberal Payoff
HIGHER ability - Conservative Payoff
HIGHER ability - Neutral Payoff
HIGHER ability - Liberal Payoff

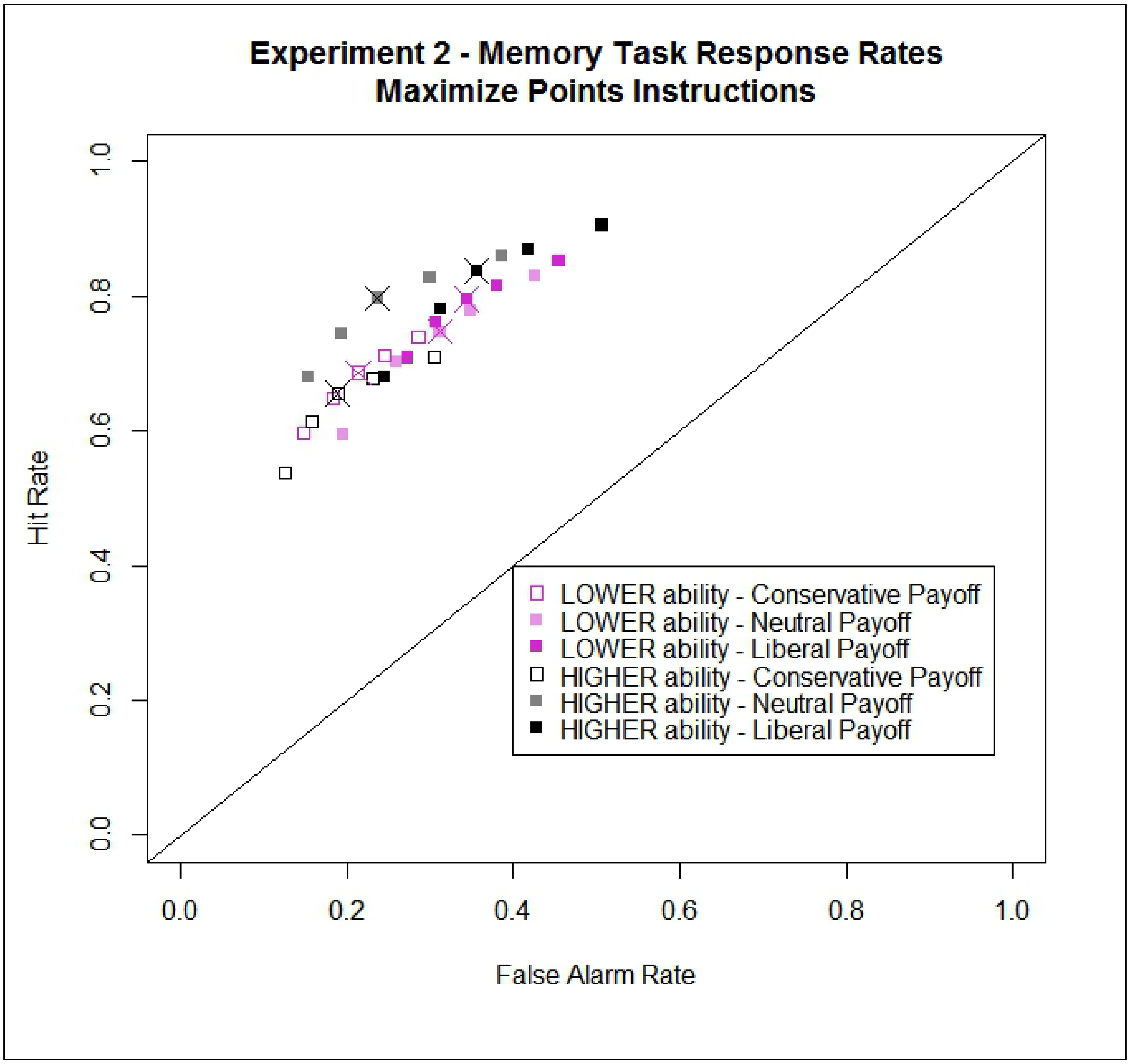


Figure 13. The ROC curves for the memory task in Experiment 2.

To visualize effects of ability on accuracy, compare the curve of the higher cognitive ability participants (black points) to the lower cognitive ability participants (colored points). According to the composite score split, higher cognitive ability participants show greater discriminability than lower cognitive ability participants, indicating that the median split was appropriate (e.g., Unsworth, 2010). To visualize the effects of condition in Figure 13, compare the curve of the liberal payoff condition (darkest colored symbols) to the curve of the neutral payoff condition (lightly shaded symbols) to the curve of the conservative payoff condition (open symbols). There should be no difference in accuracy because the payoff should only shift response bias, i.e., where along the curve the points fall. It does appear that the conservative payoff led to lower discriminability (open points are lower than the filled in points). But, more reassuringly as a manipulation check, the curves are generally pushed farther to their respective sides: the liberal curve (darkest symbols) is shifted to the right and the conservative curve (lightest symbols) to the left.

Accuracy Analyses

Full Dataset Analysis

To statistically analyze these accuracy and response bias patterns, the ROCs for each participant/condition were fit to the UVSD model. Unfortunately, like with Experiment1's reasoning task, there were many participants who did not use all of the confidence levels which made individual fits very difficult. The proposed approaches to compensate for this were to collapse across confidence levels or use a group-based slope. While collapsing across confidence levels was not feasible in Experiment 1 (a maximum of 7% of participants might potentially be fit in this way), it was feasible for Experiment 2 (up to 54% of participants potentially fit this way). So, for Experiment 2, individual fits were used where possible (i.e., for participants who used at least 3 levels of confidence) and group-based slope was used to fit the remaining participants.

Still, it is unclear why some participants failed to express many levels of confidence in their responses. Furthermore, some participants may have even used only a single response on every trial. There are two possible interpretations for why participants may have used a single response: they may have attempted to set an optimal criterion location or they may have failed to follow the instructions. For the former interpretation, using a single response for every trial, which reflects a very extreme criterion location, would not be optimal in the neutral condition. This suggests, then, that participants who used a single response may have failed to follow instructions and so their data do not inform the hypotheses. Accordingly, a cleaner test of the hypothesis includes only the participants who used multiple confidence ratings in every condition, allowing separate ROCs to be fit.

I will report analyses of data for all participants who used multiple confidence ratings in every condition using both a Full Dataset based on the combined including group-based slope fits and individual fits (all 262 participants) and Subset from just the participants' individual fits (81 participants, with 45 higher cognitive ability and 36 lower cognitive ability participants). The same median split was used in both the Full Dataset and Subset analyses.

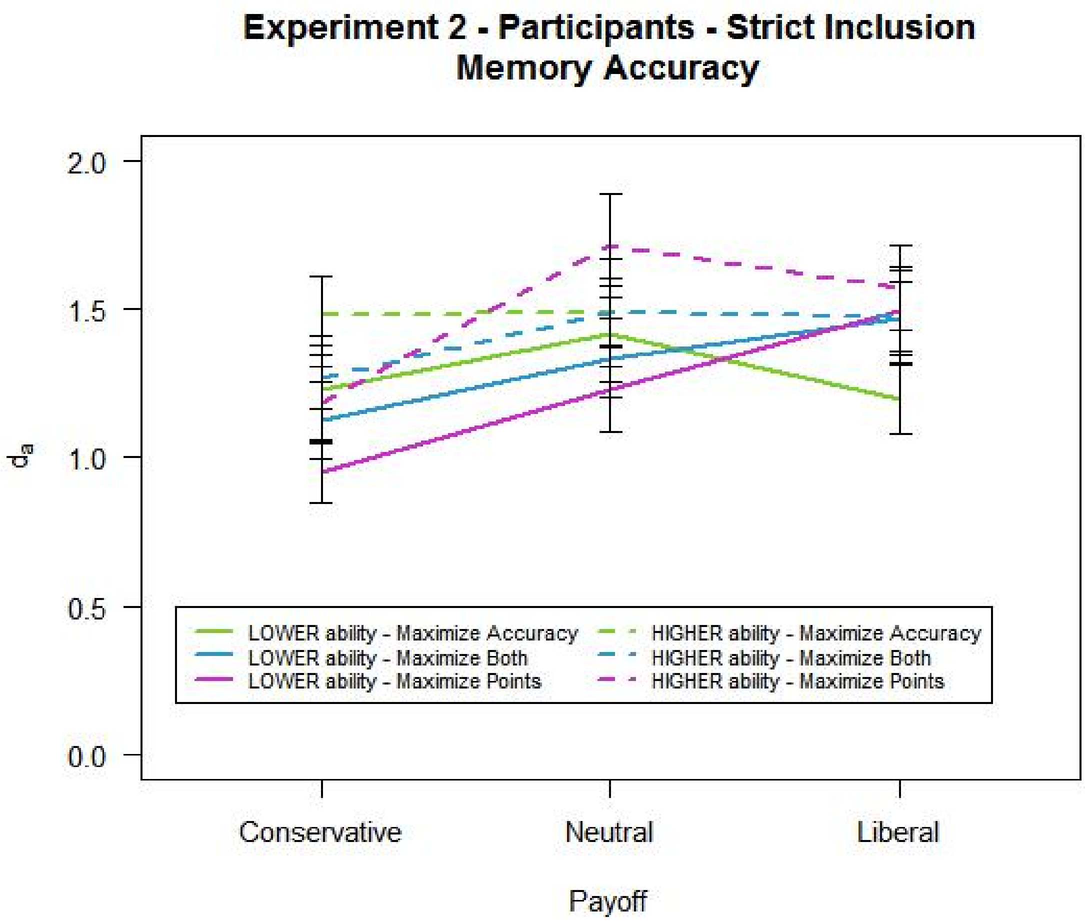


Figure 14. Plot of accuracy, measured by d_a, by ability, instruction condition, and payoff condition for all participants in Experiment 2.

To statistically analyze d_a (plotted in Figure 14), 2x3x3 mixed ANOVAs (cognitive ability by instruction condition by payoff condition) were conducted as a manipulation check to show that higher cognitive ability participants did have better discriminability than lower cognitive ability participants, indicating an appropriate median split. Figure 14 shows that higher cognitive ability participants (dotted lines) were more accurate, measured with d_a, than the lower cognitive ability participants (solid lines), $F(1, 256) = 4.34$, $p = 0.003$, $\eta^2_G = 0.01$. However, there was a significant main effect of payoff condition ($F(2, 512) = 10.02$, $p < 0.001$, $\eta^2_G = 0.01$), unexpected because the payoff condition should only affect response bias, not discriminability. The conservative payoff (far left points) led to lower accuracy than the neutral payoff ($t(261) = 3.98$, $p < 0.001$, Cohen's $d = 0.25$) and the liberal payoff ($t(261) = -3.68$, $p < 0.001$, Cohen's $d = 0.22$).

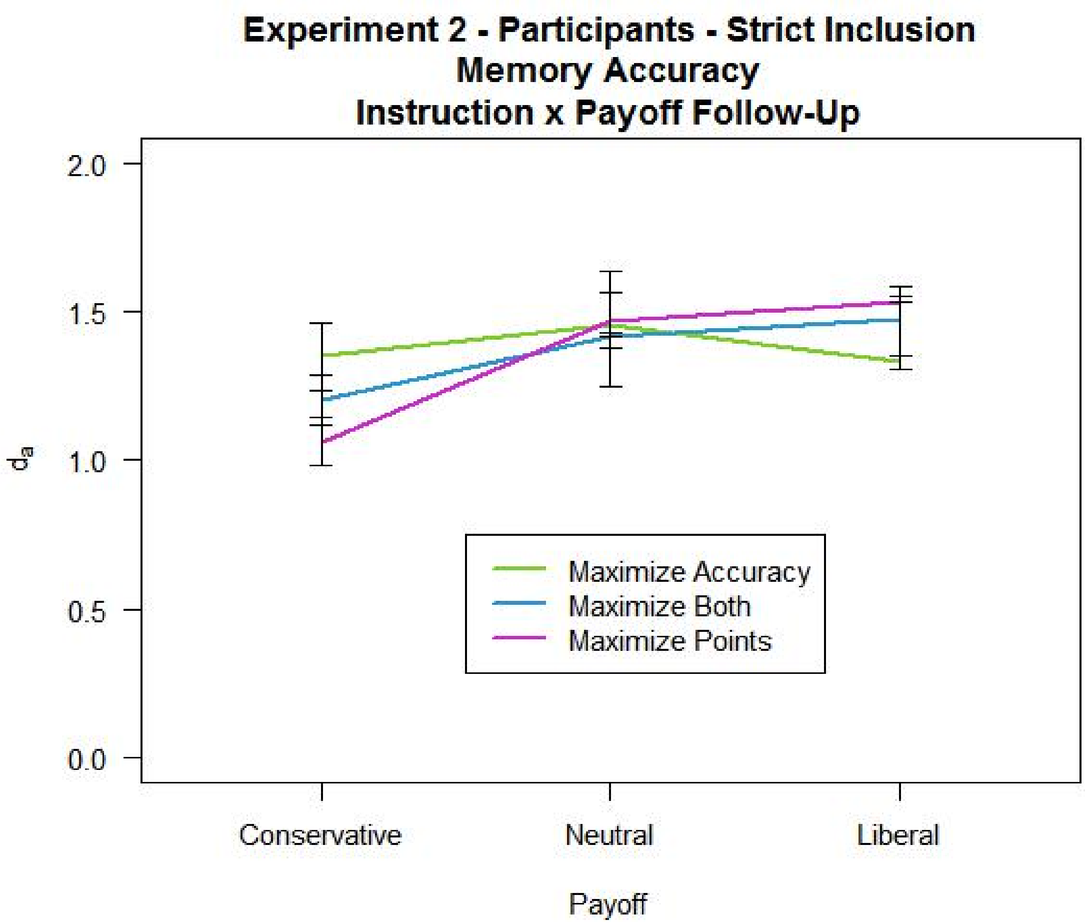


Figure 15. Plot of accuracy, measured by d_a, by instruction condition and payoff condition for all participants in Experiment 2.

There was also an unexpected significant interaction of instruction condition and payoff condition, $F(4, 512) = 2.69$, $p = 0.02$, $\eta^2_G = 0.008$. This interaction, collapsed across cognitive ability, is shown in Figure 15. It appears that accuracy *decreased* when instructions became more points-focused (purple lines) in the conservative payoff condition (left points) but *increased* when instructions became more points-focused in the liberal payoff condition (right points). While the instructions to consider points could have potentially lead participants to focus on points to the detriment of accuracy, this should not be specific to only the conservative condition. Indeed, this interaction was not seen in the “cleaner” analysis with only participants who had individually fit slopes.

Subset Analysis: Only Participants with Individually Fit Slopes

The Full Dataset analyses were repeated on the subset of participants who had individually fit slopes. Figure 16 shows only a significant main effect of payoff condition ($F(2, 150) = 4.31$, $p = 0.02$, $\eta^2_G = 0.02$). The conservative payoff (far left points) led to lower accuracy than the neutral payoff ($t(80) = -2.49$, $p = 0.01$, Cohen’s $d = 0.28$), and the neutral payoff led to lower accuracy than the liberal payoff ($t(80) = 2.02$, $p < 0.05$, Cohen’s $d = 0.22$). Unlike in the Full Dataset analysis, there was no main effect of cognitive ability, and there was no significant interaction of payoff condition and instruction condition.

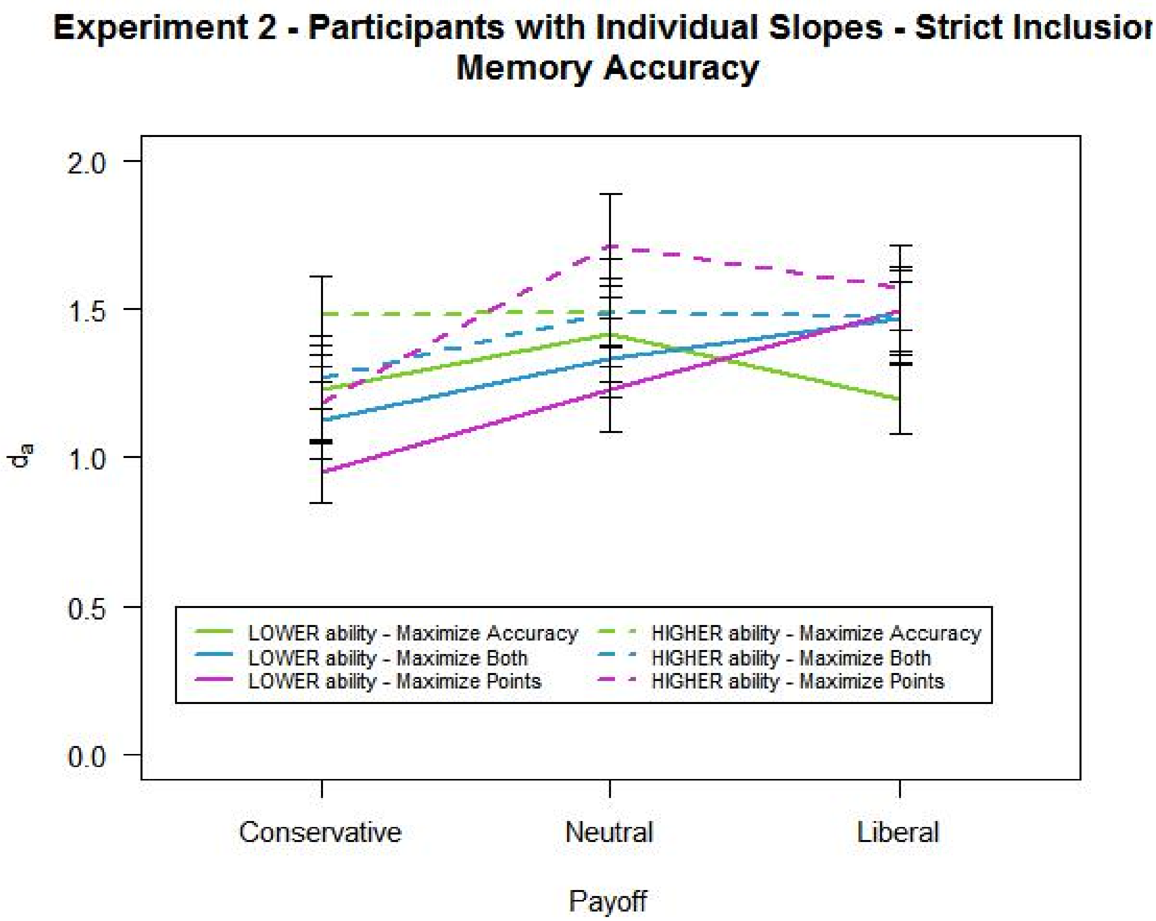


Figure 16. Plot of accuracy, measured by d_a, by ability, instruction condition, and payoff condition for only participants with individually fit slopes in Experiment 2.

While only cognitive ability effects were expected and none appeared, this sample is underpowered, with cell sizes as low as 11. Additionally, while there were no particular hypotheses of instructions effects, effects of the payoff condition were not expected. Indeed, the payoff manipulation is meant to only affect response bias. A check of this hypothesis and the other response bias hypotheses follows.

Response Bias ANOVA Analyses

Absolute Response Bias Location

Full Dataset Analysis

With indication that the median split was appropriate in the full dataset, I then analyzed the effect of cognitive ability on response bias with a 2x3x3 mixed ANOVA (cognitive ability by instructions condition by payoff condition). I hypothesized that conservative payoffs should lead to more conservative absolute response bias λ (larger values), and liberal payoffs should lead to more liberal absolute response bias λ (smaller values). I also predicted that participants would follow instructions and shift their bias more in the conservative/liberal conditions when instructed to Maximize Points.

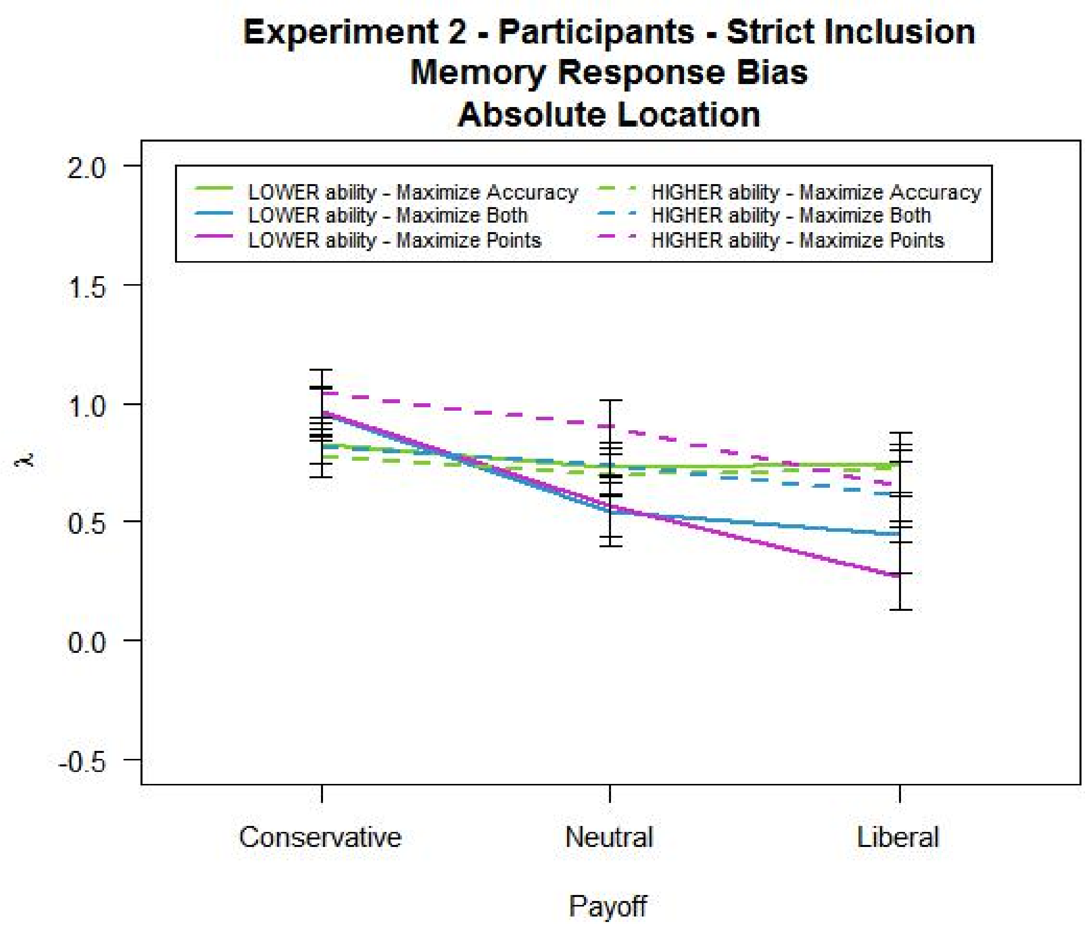


Figure 17. Plot of absolute response bias location, measured by λ, by ability, instruction condition, and payoff condition for all participants in Experiment 2.

As seen in Figure 17, there was a significant effect of payoff condition, as expected: liberal payoffs led to more liberal response bias and more conservative payoff led to more conservative response bias (a general negative slope trend), $F(2, 512) = 19.54$, $p < 0.001$, $\eta^2_G = 0.03$. There was also an interaction of instruction condition and payoff condition, $F(4, 512) = 3.67$, $p = 0.006$, $\eta^2_G = 0.01$, shown collapsed across ability level in Figure 18. As can be clearly seen by the crossover of the purple and blue lines in contrast to the green line, instructing participants to Maximize Points shifted their response bias much more than instructing them to Maximize only Accuracy. In other words, emphasizing the payoff led to greater effects of the payoff, as would be expected.

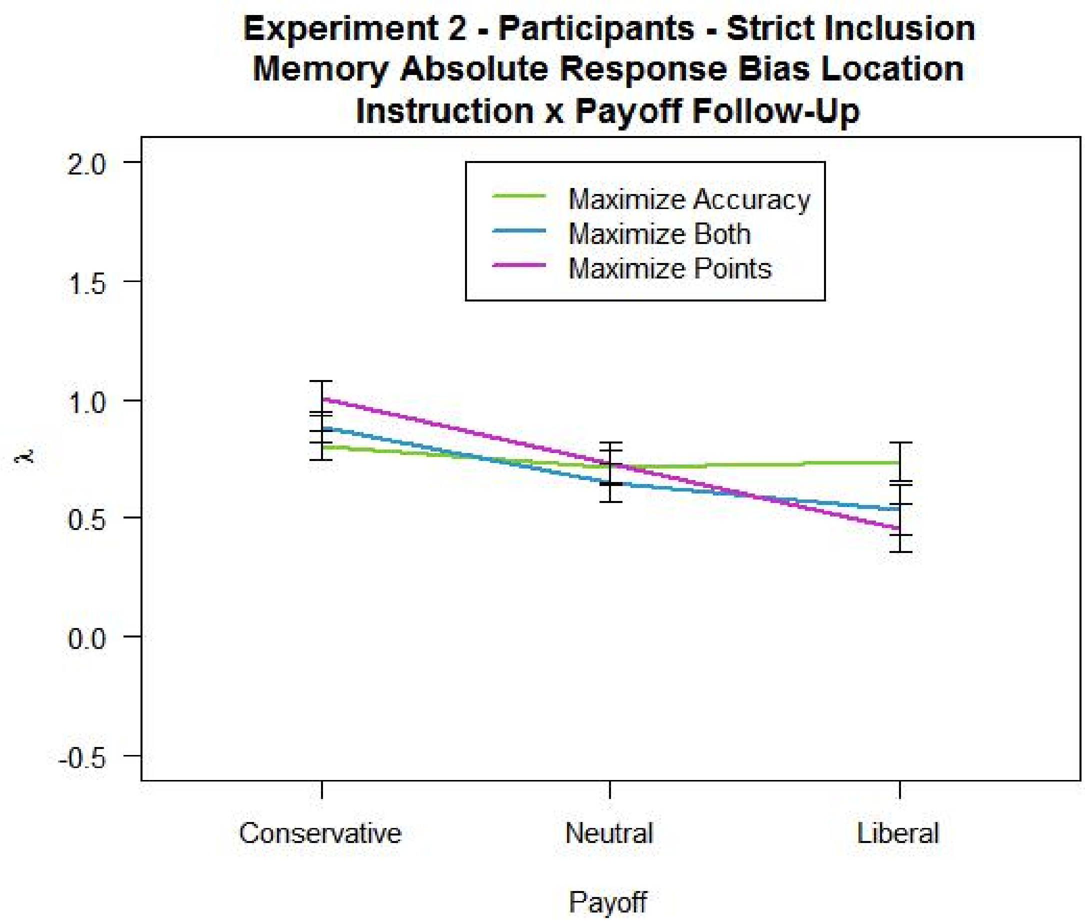


Figure 18. Plot of absolute response bias location, measured by *λ,* by instruction

condition and payoff condition for all participants in Experiment 2.

Subset Analysis: Only Participants with Individually Fit Slopes

Analogous ANOVAs on the Subset data revealed there was the same significant effect of payoff condition, where liberal payoffs led to more liberal response bias and more conservative payoff led to more conservative response bias, $F(2, 150) = 7.00$, $p = 0.001$, $\eta^2_G = 0.3$. Unlike in the Full Dataset analysis, there was no interaction of instruction condition and payoff condition.

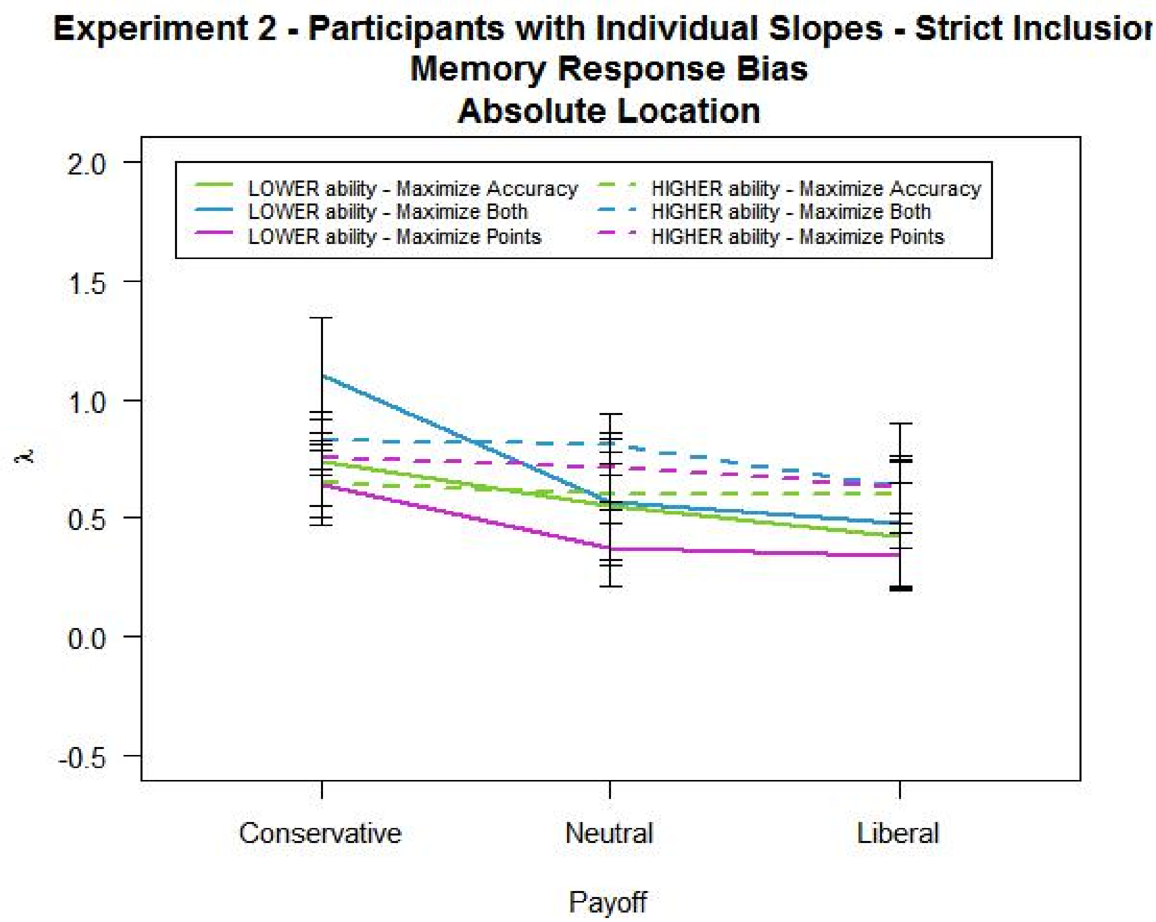


Figure 19. Plot of absolute response bias location, measured by λ, by ability, instruction condition, and payoff condition for only participants with individually fit slopes in Experiment 2.

Comparison to Optimal Response Bias Location

Full Dataset Analysis

The more relevant question for Experiment 2 was whether participants optimize their decision criteria, and whether higher cognitive ability participants are better at choosing optimal criterion locations than lower cognitive ability participants. Optimal response bias can be defined in terms of maximizing the points or in terms of maximizing accuracy; as mentioned earlier, these two goals lead to different optimal response bias locations in this experiment because of the payoff manipulation. To see effects on optimality of response bias, I conducted a 2x3x3 mixed ANOVA (cognitive ability by instructions condition by payoff condition) on two dependent variables: difference from optimal when maximizing points and difference from optimal when maximizing accuracy, defined for each payoff condition/subject. Again, I predicted a main effect payoff manipulation (conservative and liberal payoffs lead to less optimal response bias because of tension between defining optimal). I also predicted a 2-way interaction of instructions and payoff (e.g., instructions to Maximize Points leads to closer to optimal bias for points, but Maximize Accuracy does not). And finally, I predicted a 3-way interaction of cognitive ability, instructions, and payoff: higher cognitive ability participants would be better at optimizing appropriate response bias based on instructions when there is conflict in the optimal response bias (i.e., in the conservative and liberal conditions).

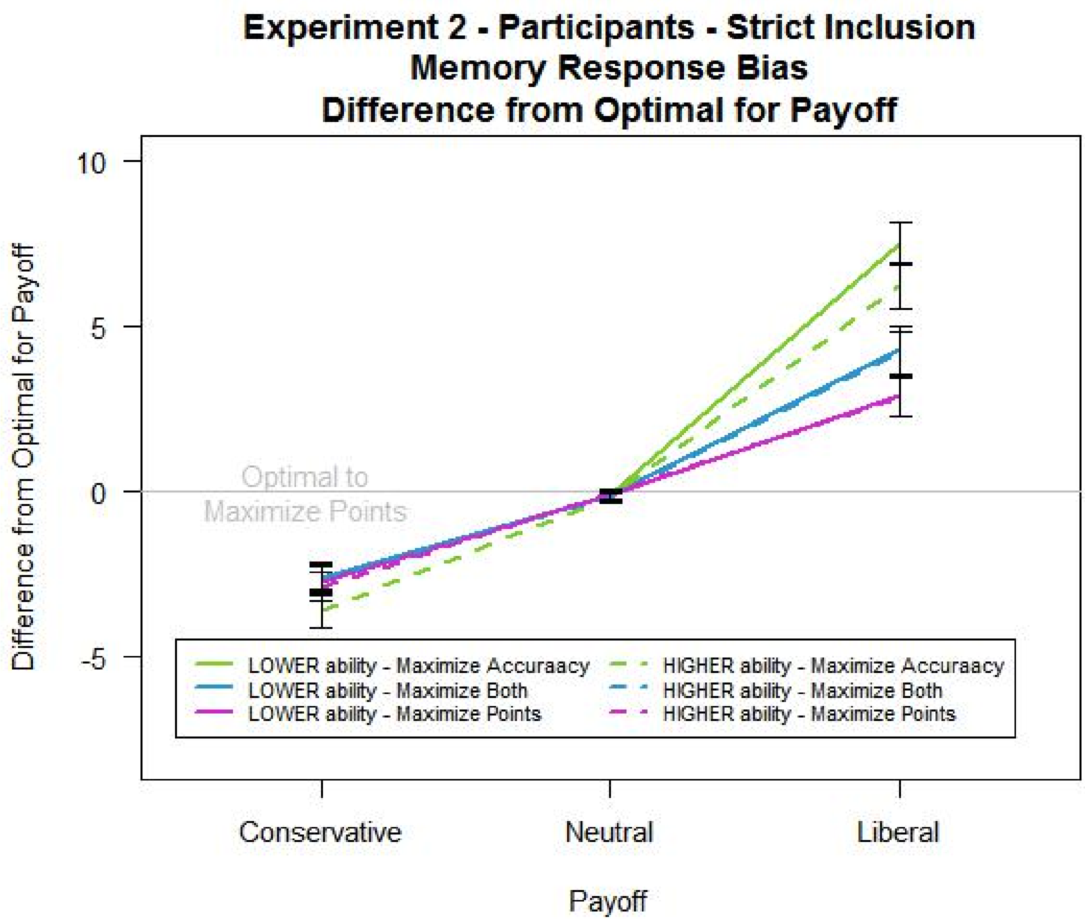


Figure 20. Plot of difference from optimal response bias location, defined in terms of optimal payoff points, by ability, instruction condition, and payoff condition for all participants in Experiment 2.

Figure 17 shows the difference from optimal for payoff: a positive number means that the participant was more conservative than optimal, and a negative number means that the participant was more liberal than optimal. As expected, there was a main effect of instructions on the difference from optimizing for points, $F(2, 256) = 10.89$, $p < 0.001$, $\eta^2_G = 0.03$: participants who were told to Maximize Accuracy (green lines) were less optimal for points (farther away from the gray Optimal line) than participants who were told to Maximize Points (purple lines), or to Maximize Both (blue lines). There was also a main effect of payoff condition, $F(2, 512) = 415.77$, $p < 0.001$, $\eta^2_G = 0.52$: the neutral payoff condition led to very close to optimal response bias (middle points very near the gray optimal line), but participants were not conservative enough in the conservative payoff condition (left points more liberal than optimal) nor liberal enough in the liberal payoff condition (right points more conservative than optimal). There was also an interaction of instruction condition and payoff condition, $F(4, 512) = 14.90$, $p < 0.001$, $\eta^2_G = 0.07$; participants in the accuracy condition were particularly non-optimal in the liberal payoff condition (green lines in the rightmost category are higher than the others). Contrary to my hypotheses, there were no effects or interactions with cognitive ability.

In looking at all participants in Experiment 2, there was no strong evidence that higher cognitive ability participants were particularly more optimal for points in the non-neutral payoff conditions when instructed to Maximize Points. In other words, higher cognitive ability participants did not maintain the goal of the instructions significantly better than lower cognitive ability participants. Looking at how participants differed from optimal response bias that maximized accuracy follows.

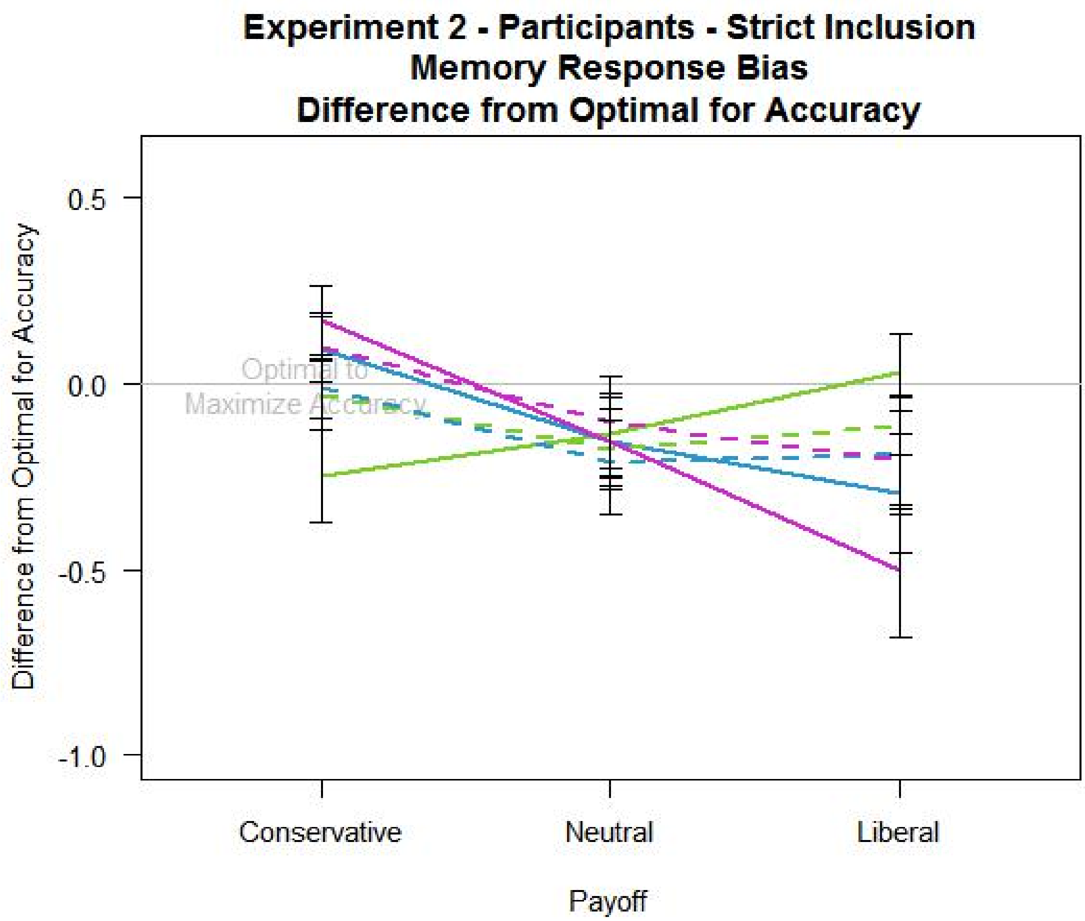


Figure 21. Plot of difference from optimal response bias location, defined in terms of optimal accuracy, by ability, instruction condition, and payoff condition for all participants in Experiment 2.

Plotted in Figure 21 is how different participants' criteria were from optimal response bias for accuracy, where a positive number is more conservative than optimal and negative is more liberal than optimal. One thing to note is that the scale for optimizing for accuracy is very different from the scale for optimizing for points. The likelihood ratio for optimizing for accuracy is around 1, but the likelihood ratio for optimizing for points is 10 in the conservative payoff and 0.1 in the liberal payoff conditions because of the particular points used here. This order of magnitude difference is reflected in the degree to which participants may need to shift in order to maximize the points they can earn, hence the much larger scale in that figure.

In analyzing these data, as expected, there was a main effect of payoff condition, $F(2, 256) = 6.38$, $p = 0.002$, $\eta^2_G = 0.02$: participants were more conservative in the conservative payoff condition and more liberal in the liberal payoff condition (a general negative slope trend in Figure 21). Indeed, this further shows that participants were influenced as predicted by the payoff condition. There was also an interaction of instruction condition and payoff condition, $F(4, 512) = 3.60$, $p = 0.006$, $\eta^2_G = 0.02$; as would be expected, participants in the Maximize Points condition were particularly non-optimal for accuracy in the conditions where they *should* be influenced by points (the left and right ends of the purple lines are farther from optimal). Contrary to my hypotheses, there were no effects or interactions with cognitive ability.

Subset Analysis: Only Participants with Individually Fit Slopes

The same analyses were repeated on the Subset of data. The only significant effect was that of payoff condition, $F(2, 150) = 203.44$, $p < 0.001$, $\eta^2_G = 0.63$: the neutral payoff condition led to very close to optimal response bias (middle points very near the gray optimal line), but participants were not conservative enough in the conservative payoff condition (left points more liberal than optimal) nor liberal enough in the liberal payoff condition (right points more conservative than optimal).

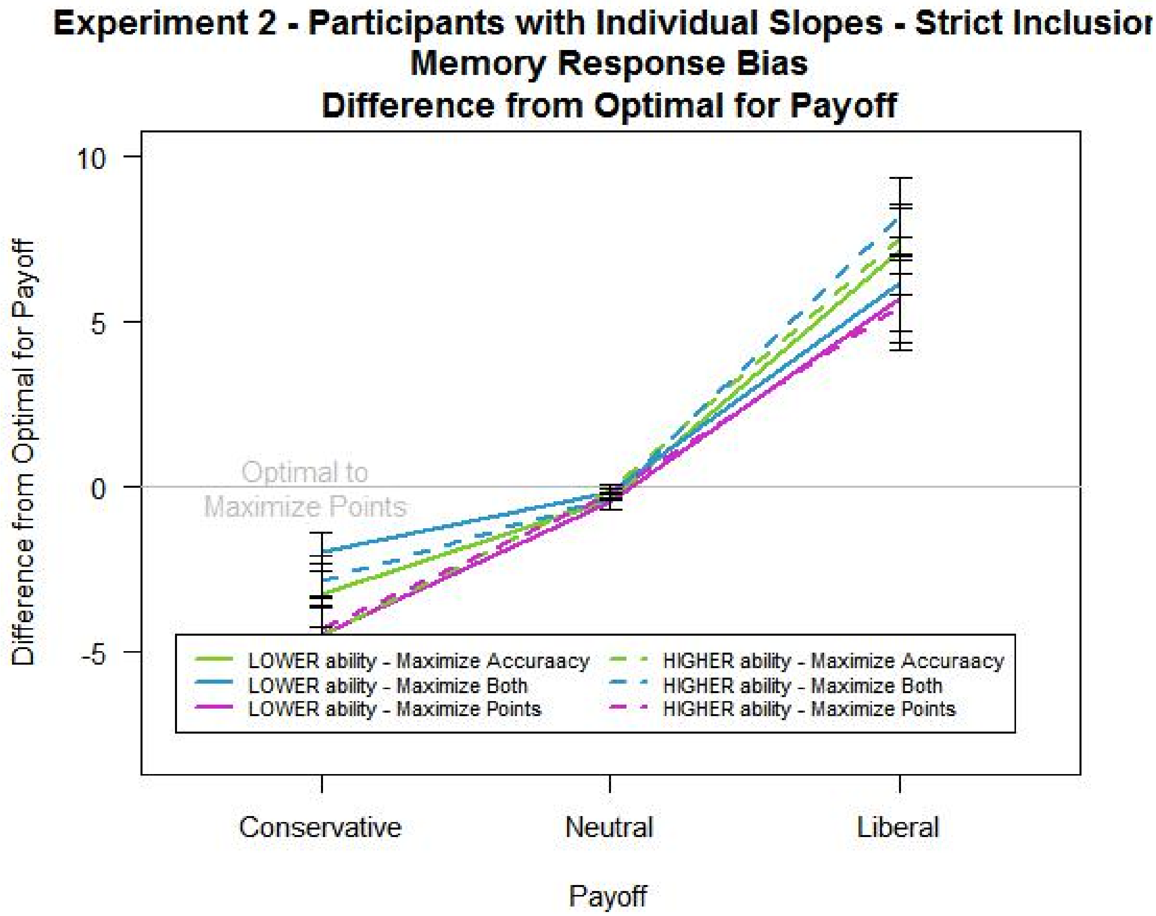


Figure 22. Plot of difference from optimal response bias location, defined in terms of optimal payoff points, by ability, instruction condition, and payoff condition for only participants with individually fit slopes in Experiment 2.

Figure 23 shows how different participants' criteria were from optimal response bias for accuracy, where a positive number is more conservative than optimal and negative is more liberal than optimal. There were no significant effects; as can be seen in the figure, the error bars are quite large.

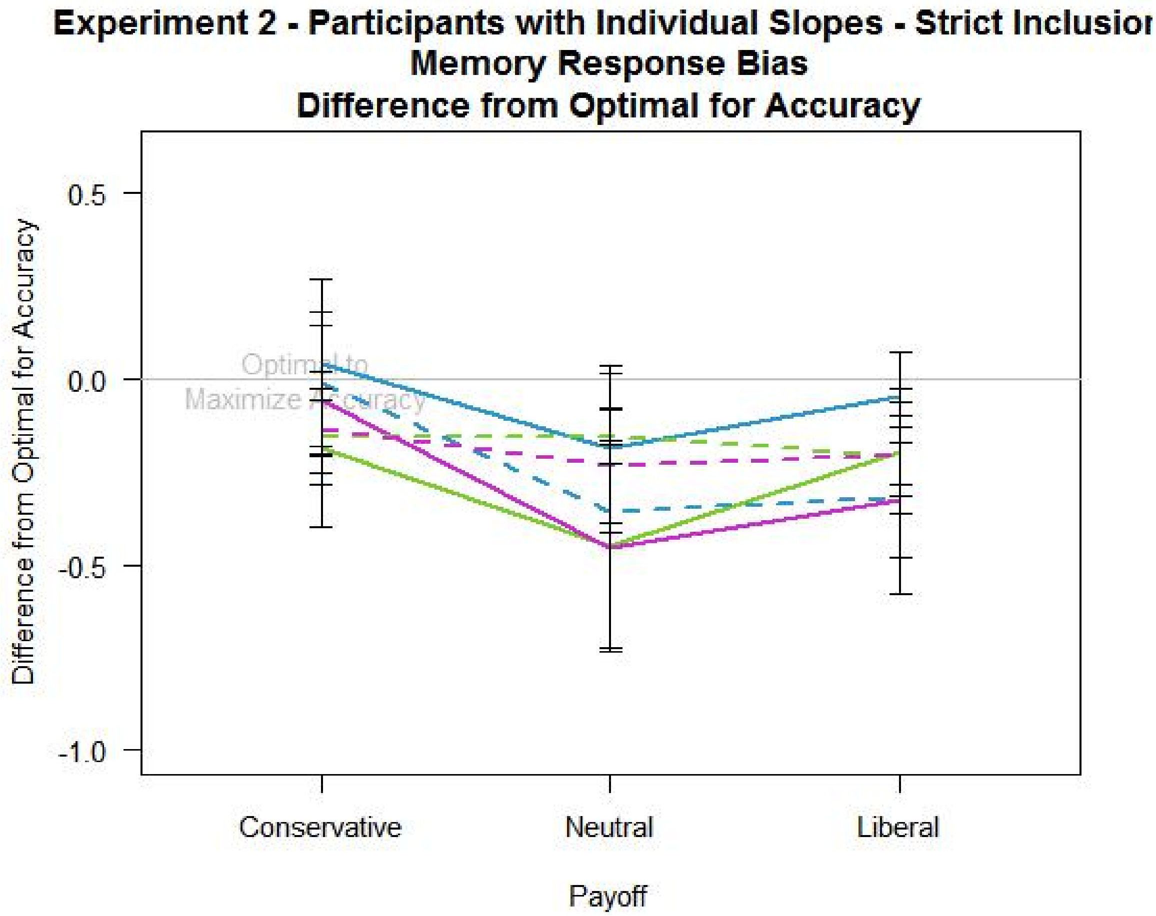


Figure 23. Plot of difference from optimal response bias location, defined in terms of optimal accuracy, by ability, instruction condition, and payoff condition in Experiment 2.

So in analyzing only the participants with individually fit slopes, it appears that both higher and lower cognitive ability participants were similar in optimizing for accuracy, regardless of instructions.

Variability in Participants' Difference from Optimal Criterion

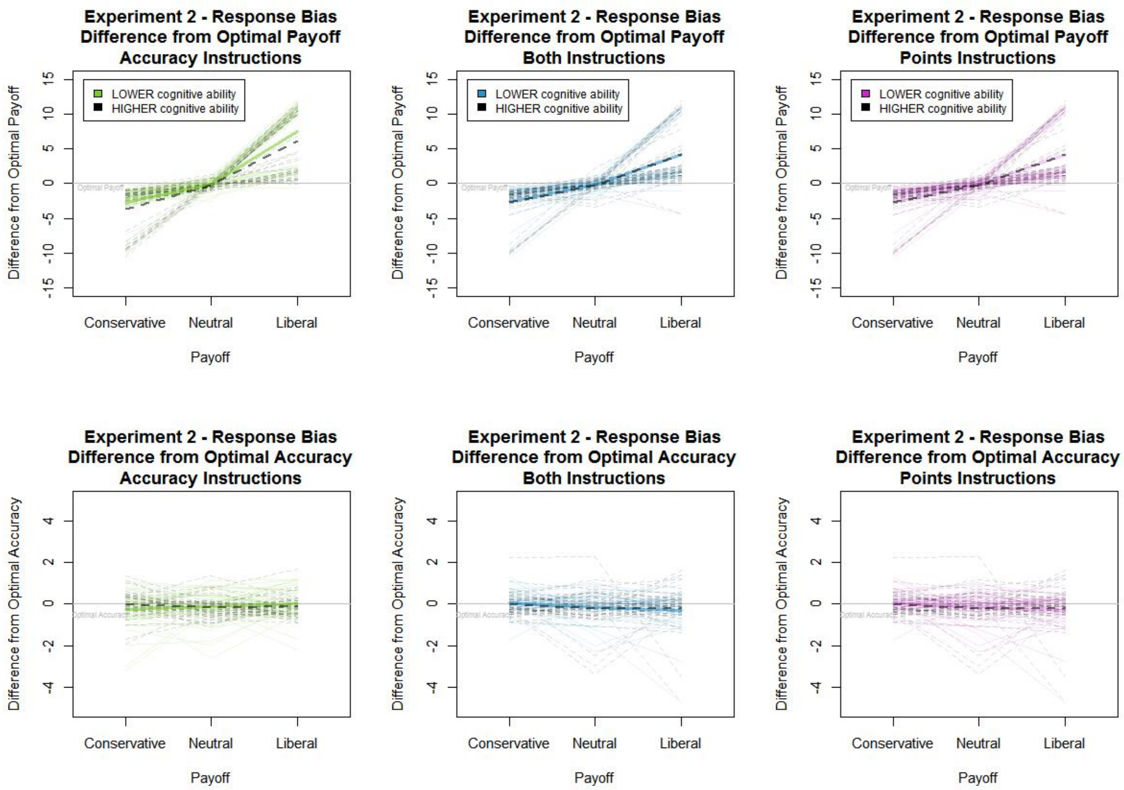


Figure 24. Plots of response bias difference from optimal as lines drawn between the payoff conditions in Experiment 2. The top three are in comparison to optimizing payoff points and the bottom three are in comparison to optimizing accuracy.

Figure 24 shows all participants' difference from optimal response bias where the closer to the horizontal 0 line, the more optimal. Lower cognitive ability participants are represented with colored lines and higher cognitive ability participants are in dotted black lines. It is clear that participants vary greatly in how close to optimal they are when setting their decision criteria. Indeed, in the top row of the figure, there are a subset of participants who clearly fail to approach optimal response bias based on the payoff in the conservative and neutral conditions (their lines extend to -10 or 10). However, cognitive ability as measured in the current study does not appear to predict this variability.

Experiment 2 Discussion

While the Full Dataset was analyzed, it consisted of a combination of model fitting approaches because almost half of the participants used too few confidence levels to be individually fit. Accordingly, a Subset of participants whose data were able to be individually fit was also analyzed. Examining effect on accuracy showed that higher cognitive ability participants did perform better than lower cognitive ability participants in the Full Dataset (though not in the underpowered Subset), but surprisingly, there was a significant effect of payoff in both the Full Dataset and the Subset: the conservative payoff led to lower discriminability even though payoff condition was meant only to affect response bias.

Payoff condition had its expected effect on response bias: participants shifted their absolute criterion location to be more conservative when in the conservative payoff condition and to be more liberal in the liberal payoff condition, and indeed, instructing participants to pay Maximize Points led to greater shifts than the other instructions.

However, there were no effects of cognitive ability on response bias location, nor on optimizing response bias location. So, contrary to my hypotheses, there was no strong evidence that higher cognitive ability participants are better able to choose optimal response biases than lower cognitive ability participants.

CHAPTER 4

GENERAL DISCUSSION

Conclusions from the Current Study

The overarching goal of the current study was to answer whether differences in response bias in memory and reasoning can be explained by cognitive ability and how these might affect bias manipulations in experimental design. Experiments 1 and 2 showed that in general, individual differences in response bias tendency in a recognition memory task can not be explained by cognitive ability as measured in this study. There was evidence in Experiment 1 to indicate that cognitive ability does play a role in a belief bias reasoning task, however. Specifically, higher cognitive ability participants tended to be more conservative in their response bias than lower cognitive ability participants in reasoning. Experiment 2 directly examined the extent to which this increased shifting was related to optimal responding. Contrary to predictions, higher cognitive ability participants did not use more optimal response bias than lower cognitive ability participants.

Limitations and Strengths

There are some limitations to the current study. First, the necessity of large samples to test for potentially small effects required use of an online sample from Mechanical Turk. While some researchers have argued that this population is equally generalizable as the more typical undergraduate population (e.g., Paolacci, Chandler, & Ipeirotis, 2010), there were some issues to note in the present study. For example, because SAT or ACT test score was a measure of cognitive ability, online participants were screened based on whether they could report a score within the plausible range of scores and years given the age restriction. While participants could earn money for completing the screening, there may have been particular motivation to access the rest of the study; participants were sometimes caught misreporting their scores, but it was nearly impossible to determine which scores were intentionally misleading and which were misremembered. While the self-report aspect of test scores was also required in the lab sample, participants may not have been as motivated to misreport their scores because they were already eligible for participation by the time they came to the lab. Their test scores may have also been more salient to them, having needed to report them for entrance to the university, than the online sample who may have forgotten their precise scores. Additionally, the poor performance in the Go/No-Go task in particular suggests differences in motivation in the online sample. Also, the lack of control of environment in the online sample in comparison to the lab sample is a general concern. It is unclear how many participants had distractions in the online sample and it is almost certain that they had a different range of distractions than in the lab sample. Still, the findings were generally qualitatively similar across the online and lab sample in Experiment 1, indicating that the differences in sample did not appear to affect this particular study.

Another issue was that in Experiment 2, participants may have chosen a strategy that would undermine the goal of the experiment to measure response bias because of the lack of penalties in the payoffs. In particular, by instructing some participants to Maximize Points, but without penalizing incorrect answers, participants in, say, the liberal payoff condition might adopt the strategy of only responding that items were old. Because of how disparate the payoff was (1 point vs. 10 points), participants would have to weigh 10 correct rejections vs. 1 hit in this liberal condition. That is, while earning 1 point for a correct rejection is better than earning 0 points for a false alarm if the item is a lure, with an equal number of targets and lures, a participant might instead choose to adopt the strategy of always saying that the item is a target because the resulting expected value is higher. In this case, participants do indeed shift their bias towards a more liberal one, but the task is no longer testing memory. An obvious solution would be to include penalties for incorrect answers. This is usually how payoff manipulations are implemented, though they were not implemented that way in the current study in order for clarity and simplicity for the online participants.

Another limitation to the study came from the difficulty of fitting the signal detection model to the behavioral data. Because a large proportion of participants did not distribute their responses across all of the confidence levels, alternative approaches to measuring accuracy and response bias were needed. In Experiment 1, the reasoning task was estimated with the equal-variance signal detection model because there were too few trials to fit the unequal-variance signal detection model. While recent research indicates that this might be appropriate (Trippas et al., 2016), that research is as yet not peer-reviewed nor published. The other approach, implemented in Experiment 2, used a group-based slope for roughly 2/3 of the participants in at least one condition, but this requires the assumption that such a slope is a good estimate for each individual's true slope. This assumption may not be appropriate, especially in a study of individual differences. Easy remedies for future research include presenting more trials and instructing participants to use all confidence levels, but the latter does pose other potential complications. Specifically, instructing participants to use all confidence levels would obscure participants' natural response bias, though it could provide more validity in comparison to other studies that instruct their participants to do so.

Despite these limitations, this study still had many strengths. Experiment 1 was the first to explore response bias in recognition memory and reasoning simultaneously. While there are common processes in both recognition and reasoning (Heit et al., 2012), they have often been studied independently. Experiment 2 also provided a thorough examination of the tension between experiment demands (such as the payoff conditions) and overarching goal maintenance (instructions to maximize accuracy or points) and how cognitive ability might explain how that tension is resolved.

Future Directions

Though cognitive ability was not found to predict individual differences in response bias in the current study, further research is still needed. For example, it is still unclear what might predict how participants choose to compromise their response bias when faced with conflicting optimal options. Some participants may be better at choosing a compromise that optimizes both, or some may be better at maintaining goals throughout the task. Further research could explore what might predict this. Additionally, the current study used the Go/No-Go task, Number-Letter task, SAT/ACT test scores, and the Cognitive Reflection Task to measure cognitive ability. A next step could be to determine whether simpler, shorter, or more measurement error-resistant measures of cognitive ability could predict response bias. Indeed, while the current tasks were used in part because a working memory capacity task was not feasible for online implementation, further simplicity of a task could increase the likelihood of use by other researchers in accounting for the “extra noise” in response bias. That is, a simpler measure would be more likely to be adopted as a potential covariate of response bias tendency or optimality in future research which would reduce the “noise” of individual differences in response bias on parameter estimation or experimental manipulation.

Additionally, the current experiment examined signal detection-based measures of response bias, particularly at the binary decision point (e.g., of deciding between old/new in the memory task). Future research could examine how individual differences in response bias might be reflected in the full ROC/full scale. For example, certain participants may shift all of their criteria, or maybe only some of the criteria. Some participants might exhibit a response bias towards using the *extreme* ends of the scale rather than distributing their responses across the full scale, as has been found in marketing and survey contexts (e.g., Chami-Castaldi, Reynolds, & Wallace, 2008; Greenleaf, 1992; Johnson, 2003; Naemi et al., 2009). Indeed, while sometimes treated as a continuous scale, confidence scales are Likert scales and so other research on Likert scales may be applicable in examining response bias here; other response bias styles than extreme response bias may also be relevant (e.g., Baumgartner & Steenkamp, 2001; R. Jones & Rorer, 1973).

Furthermore, response bias is sometimes measured not on a *global* scale (across the whole list) but on a *local* scale (e.g. Benjamin, Diaz, & Wee, 2009; Jones et al., 2015; Treisman & Faulkner, 1984; Treisman & Williams, 1984). That is, sequential dependencies, or what happened on previous trials such as a string of targets in a row, may affect the short-term adjustments participants make during a test list. It could be important to know what predicts the extent to which participants are influenced by sequential dependencies and adjust their criteria accordingly.

Relatedly, metacognition and its role in response bias could be predicted by cognitive ability. Indeed, a study on metacognition found that the use of confidence ratings scales was stable across lists (Thompson & Mason, 1996). Additionally, the metacognitive processes of long-term and short-term adjustments to response criteria could reflect the ability to maintain and/or update these goals. Other potentially relevant metacognitive processes include processing one's own uncertainty (de Gardelle & Mamassian, 2014), or how that processing could lead to better metacognitive sensitivity, i.e., greater confidence in a response maps onto better discriminability (Maniscalco & Lau, 2012). More recent work by Jonker (2016) indicates that metacognition of confidence in a source memory task can predict response bias in recognition and recall. An interesting question could therefore be: what metacognitive processes separate from other cognitive ability processes influence response bias?

General Conclusions

That there are individual differences in response bias can be seen quite clearly by the variability in response bias location and response bias shifting. Yet, as measured in the current study, it does not appear that cognitive ability predicts response bias in memory. In reasoning, though, higher cognitive ability participants tend to be more conservative in responding. Further exploration of what predicts the variability in response bias is important because it could be used to reduce the error variance in statistical analyses, and therefore provide more power to the experimental design and analyses. Indeed, when those designs include bias manipulations, such as payoffs, the impact on response bias may not be entirely monolithic: some participants in some payoff conditions respond quite differently. As seen here, bias manipulations are generally effective, but they may also introduce unaccounted for noise when one ignores the interaction of individual differences on such manipulations. Thus, while the current study focused on cognitive ability as a predictor, understanding individual differences in response bias in memory and reasoning more generally ultimately enriches the understanding and theories of cognition.

BIBLIOGRAPHY

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., … Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–30. doi:10.3758/s13421-012-0204-6

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*(5), 724–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9337590

Baddeley, A. D. (1996). Exploring the Central Executive. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 5–28. doi:10.1080/713755608

Baddeley, A. D., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation*, *8*, 47–89. doi:10.1016/S0079-7421(08)60452-1

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*(2), 81–99.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*(4), 548–564. doi:10.1037//0033-295X.97.4.548

Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological Review*, *116*(1), 84–115. doi:10.1037/a0014351.Signal

Bohil, C., & Maddox, W. (2001). Category discriminability, base-rate, and payoff effects in perceptual categorization. *Perception & Psychophysics*, *63*(2), 361–376. Retrieved from http://link.springer.com/article/10.3758/BF03194476

Brandt, J., Corwin, J., & Krafft, L. (1992). Is verbal recognition memory really different in Huntington's and Alzheimer's disease? *Journal of Clinical and Experimental Neuropsychology*, *14*(5), 773–784.

Brébion, G., Smith, M. J., Amador, X., Malaspina, D., & Gorman, J. M. (1998). Word recognition, discrimination accuracy, and decision bias in schizophrenia: association with positive symptomatology and depressive symptomatology. *The Journal of Nervous and Mental Disease*, *186*(10), 604–9.

Cacioppo, J., & Petty, R. (1982). The need for cognition. *Journal of Personality and Social …*, *42*(1), 116–131. doi:10.1037//0022-3514.42.1.116

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333. doi:10.1037//0022-3514.67.2.319

Cattell, R. (1994). Theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, *54*(1), 1–22.

Chami-Castaldi, E., Reynolds, N., & Wallace, J. (2008). Individualised rating-scale procedure: A means of reducing response style contamination in survey data? *The Electronic Journal of Business Research Methods, 6*(1), 9–20.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review*, *15*(5), 906–26. doi:10.3758/PBR.15.5.906

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*(4), 692–712. doi:10.3758/PBR.15.4.692

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. doi:10.1080/14640748108400805 de Gardelle, V., & Mamassian, P. (2014). Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychological Science*. doi:10.1177/0956797614528956

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22.

Del Missier, F., Mäntylä, T., Hansson, P., Bruine de Bruin, W., Parker, A. M., & Nilsson, L.-G. (2013). The multifold relationship between memory and decision making: an individual-differences study. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1344–64. doi:10.1037/a0032379

Dowens, M., & Calvo, M. (2003). Genuine memory bias versus response bias in anxiety. *Cognition & Emotion*, *17*(6), 843–857. doi:10.1080/02699930244000381

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*(3), 831–63. doi:10.1037/a0019634

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: a reply to Klauer and Kellen (2011). *Psychological Review*, *118*(1), 155–63. doi:10.1037/a0021774

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. University Microfilms.

Engle, R., & Tuholski, S. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of …*, *128*(3), 309–331. doi:10.1037//0096-3445.128.3.309

Evans, J. S. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*(3), 295–306. doi:10.3758/BF03196976

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. doi:10.1257/089533005775196732

Frenkel, T. I., Lamy, D., Algom, D., & Bar-Haim, Y. (2009). Individual differences in perceptual sensitivity and response bias in anxiety: Evidence from emotional faces. *Cognition & Emotion*, *23*(4), 688–700. doi:10.1080/02699930802076893

Friedman, N. P., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P., & Hewitt, J. K. (2009). Individual differences are almost entirely genetic in origin. *Journal of Experimental Psychology*, *137*(2), 201–225. doi:10.1037/0096-3445.137.2.201.Individual

Gilinsky, A. S., & Judd, B. B. (1994). Working memory and bias in reasoning across the life span. *Psychology and Aging*, *9*(3), 356–71.

Gillespie, C., & Eysenck, M. (1980). Effects of introversion–extraversion on continuous recognition memory. *Bulletin of the Psychonomic Society*, *15*(4), 233–235.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition*. doi:10.3758/BF03198438

Graham, L. M. (2007). Need for cognition and false memory in the Deese–Roediger–McDermott paradigm. *Personality and Individual Differences*, *42*(3), 409–418. doi:10.1016/j.paid.2006.07.012

Greenleaf, E. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328–351. Retrieved from http://poq.oxfordjournals.org/content/56/3/328.short

Handley, S. J., & Trippas, D. (2015). Dual processes, knowledge, and structure: A critical evaluation of the Default Interventionist account of biases in reasoning and judgment. In *The Psychology of Learning and Motivation* (pp. 1–32).

Heim, A. W. (1970). *AH4 Group Test of General Intelligence*. London: National Foundation for Educational Research.

Heit, E., & Hayes, B. K. (2011). Predicting reasoning from memory. *Journal of Experimental Psychology: General*, *140*(1), 76–101. doi:10.1037/a0021488

Heit, E., Rotello, C. M., & Hayes, B. K. (2012). Relations between memory and reasoning. In *The psychology of learning and …* (Vol. 57, pp. 57–101). doi:10.1016/B978-0-12-394293-7.00002-9

Hirshman, E. (1995). Decision processes in recognition memory: criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 302–313. doi:10.1037//0278-7393.21.2.302

Huh, T. J., Kramer, J. H., Gazzaley, A., & Delis, D. C. (2006). Response bias and aging on a recognition memory task. *Journal of the International Neuropsychological Society : JINS*, *12*(1), 1–7. doi:10.1017/S1355617706060024

Johnson, T. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563–583.

Jones, P. R., Moore, D. R., Shub, D. E., & Amitay, S. (2015). The role of response bias in perceptual learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(5), 156–1470.

Jones, R., & Rorer, L. (1973). Response biases and trait descriptive adjectives. *Multivariate Behavioral Research*, *8*(3), 313–330.

Jonker, T. R. (2016). Individual Differences in Incorrect Responding and the Retrieval Individual Differences in Incorrect Responding and the Ability to Discriminate the Source of the Products of Retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*, 1163–1177. doi:10.3758/s13421-012-0226-0

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, (2012), 1–26. doi:10.3758/s13423-014-0608-3

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*(4), 465–78. doi:10.3758/PBR.17.4.465

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*(4), 560–572.

Logan, G. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. Retrieved from http://psycnet.apa.org/psycinfo/1994-97487-005

Macmillan, N. A., & Creelman, C. (2005). *Detection Theory: A User's Guide*. (2nd, Ed.)*Detection Theory A users guide* (Vol. Standort:). Psychology Press.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*(3), 401–413. doi:10.1037/0033-2909.107.3.401

Maddox, W. T., & Bohil, C. J. (2004). Probability matching, accuracy maximization, and a test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics*, *66*(1), 104–118. doi:10.3758/BF03194865

Malmberg, K. J. (2008). Toward an understanding of individual differences in episodic memory: Modeling the dynamics of recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (Vol. 48, pp. 313–349). San Diego: Elsevier Academic Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–30. doi:10.1016/j.concog.2011.09.021

Mickes, L., Wixted, J., & Wais, P. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858–865.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, a H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. doi:10.1006/cogp.1999.0734

Myung, I. J. (2000). The Importance of Complexity in Model Selection. *Journal of Mathematical Psychology*, *44*(1), 190–204. doi:10.1006/jmps.1999.1283

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261–86. doi:10.1111/j.1467-6494.2008.00545.x

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *57*(1), 33–60. doi:10.1080/02724980343000116

Oakhill, J., Johnson-Laird, P., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, *31*, 117–140.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. doi:10.2139/ssrn.1626226

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–203. doi:10.1037/a0033044

Pitt, M. a, Myung, J. I., Montenegro, M., & Pooley, J. (2008). Measuring model flexibility with parameter space partitioning: an introduction and application example. *Cognitive Science*, *32*(8), 1285–303. doi:10.1080/03640210802477534

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–35.

Redick, T. S., Calvo, A., Gay, C. E., & Engle, R. W. (2011). Working memory capacity and go/no-go task performance: selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(2), 308–24. doi:10.1037/a0022216

Riefer, D. M., Hu, X., & Batchelder, W. H. (1994). Response strategies in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 680–693. doi:10.1037//0278-7393.20.3.680

Roberts, R. J., Hager, L. D., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General*, *123*(4), 374–393. doi:10.1037/0096-3445.123.4.374

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Rotello, C. M., & Macmillan, N. A. (2008). Response bias in recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (Vol. 48, pp. 61–94). San Diego: Elsevier Academic Press.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. doi:10.1080/01621459.1993.10476408

Sá, W., West, R., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *91*(3), 497–510. doi:10.1037/0022-0663.91.3.497

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*(5), 1178–1197. doi:10.1037//0022-3514.83.5.1178

Simmonds, D. J., Pekar, J. J., & Mostofsky, S. H. (2008). Meta-analysis of Go/No-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. *Neuropsychologia*, *46*(1), 224–232. doi:10.1016/j.neuropsychologia.2007.07.015

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. doi:10.1037//0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–95. doi:10.1037/0022-3514.94.4.672

Stanovich, K. E., West, R., & Toplak, M. (2010). Individual differences as essential components of heuristics and biases research. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The Science of Reason: A Festschrift for Jonathan St B.T. Evans* (pp. 355–396).

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*(6), 1379–1396. doi:10.1037/0278-7393.24.6.1379

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Stupple, E., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*(1), 54–77. doi:10.1080/13546783.2012.735622

Stupple, E., Ball, L. J., Evans, J. S. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, *23*(8), 931–941. doi:10.1080/20445911.2011.589381

Thompson, W. B., & Mason, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition*, *24*(2), 226–234. doi:10.3758/BF03200883

Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 119–139. doi:10.1037//0096-1523.10.1.119

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*(1), 68–111. doi:10.1037//0033-295X.91.1.68

Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1393–402. doi:10.1037/a0032398

Trippas, D., Handley, S. J., & Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: new indices for old effects. *Frontiers in Psychology*, *5*, 631. doi:10.3389/fpsyg.2014.00631

Trippas, D., Kellen, D., Pennycook, G., Fugelsang, J., & Koehler, D. (2016). Signal detection theory and belief bias: Testing the equal variance assumption. In *International Conference on Thinking*. Providence, RI.

Trippas, D., Verde, M. F., & Handley, S. J. (2014). Using forced choice to test belief bias in syllogistic reasoning. *Cognition*, *133*(3), 586–600. doi:10.1016/j.cognition.2014.08.009

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica*, *134*(1), 16–28. doi:10.1016/j.actpsy.2009.11.010

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. doi:10.3758/BF03192720

Van Zandt, T. (2000). ROC curves and confidence judgements in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582–600.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. doi:10.1037//0022-3514.54.6.1063

Watson, J. M., Bunting, M. F., Poole, B. J., & Conway, A. R. a. (2005). Individual differences in susceptibility to false memory in the Deese-Roediger-McDermott paradigm. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 76–85. doi:10.1037/0278-7393.31.1.76

Woodard, J. L., Axelrod, B. N., Mordecai, K. L., & Shannon, K. D. (2004). Value of signal detection theory indexes for Wechsler Memory Scale-III recognition measures. *Journal of Clinical and Experimental Neuropsychology*, *26*(4), 577–86. doi:10.1080/13803390490496614

Zhu, B., Chen, C., Loftus, E. F., Lin, C., He, Q., Chen, C., … Dong, Q. (2010). Individual differences in false memory from misinformation: cognitive factors. *Memory*, *18*(5), 543–55. doi:10.1080/09658211.2010.487051

Zhu, B., Chen, C., Loftus, E. F., Lin, C., He, Q., Chen, C., … Dong, Q. (2010). Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual Differences*, *48*(8), 889–894. doi:10.1016/j.paid.2010.02.016